Ground-truth-based trajectory evaluation in videos

by

Tahir Habib Nawaz

BE in Mechatronics Engineering 2005

MSc in Vision and Robotics 2009

A dissertation submitted to

The School of Electronic Engineering and Computer Science

in partial fulfilment of the requirements for the Degree of

Doctor of Philosophy

in the subject of

Interactive and Cognitive Environments

Queen Mary University of London

Mile End Road

E1 4NS, London, UK

May 2014

Acknowledgements

This PhD Thesis has been developed in the framework of, and according to, the rules of the Erasmus Mundus Joint Doctorate on Interactive and Cognitive Environments EMJD ICE [FPA n° 2010-0012] with the cooperation of the following Universities:

Alpen-Adria-Universität Klagenfurt – AAU

Queen Mary, University of London – QMUL

Technische Universiteit Eindhoven – TU/e

Università degli Studi di Genova – UNIGE

Universitat Politècnica Catalunya – UPC

According to ICE regulations, the Italian PhD title has also been awarded by the Università degli Studi di Genova.

Acknowledgements

I would like to first thank Almighty Allah for all the blessings and successes bestowed in my life.

I am grateful to my parents and siblings for their endless love and support that have enabled me to chase my goals in life.

I would like to express my gratitude to Professor Andrea Cavallaro (my primary supervisor) and Professor Bernhard Rinner (my secondary supervisor) for their guidance, support and valuable suggestions and advices, which helped me to perform this research.

I am thankful to all the colleagues for the great time spent together and the discussions during the course of my PhD studies.

I, Tahir Habib Nawaz, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Tahir Habib Nawaz

Date: 30/10/2014

First supervisor | Second supervisor | Author

Professor Andrea Cavallaro | **Professor Bernhard Rinner** | **Tahir Habib Nawaz**

Ground-truth-based trajectory evaluation in videos

Abstract

Video tracking involves estimating the state(s) of target(s) over time on the image plane, where the sequence of target states is termed as a trajectory. Trajectory evaluation refers to the assessment of a tracker's results that may be based on quantification of the discrepancy in the estimated states with respect to the corresponding ground-truth states. In this thesis, after presenting a review of the related work, we make the following proposals for the ground-truth-based trajectory evaluation in videos.

We propose three overlap-based measures that account for the key aspects of multi-target tracking evaluation including accuracy, cardinality error and ID changes. The measures quantify tracking performance by combining accuracy and cardinality errors at frame level, computing the sequence-level tracking accuracy at varying accuracy levels, and measuring ID changes while considering the length of the track in which they occur. An extensive experimental validation of the proposed measures is conducted using four state-of-the-art multi-target trackers on challenging real-world publicly-available datasets. The proposed measures show advantages (because they are parameter independent and numerically bounded) over the existing measures and enable a thorough evaluation of trackers while identifying their strengths and weaknesses.

We present a protocol that is composed of a set of trials that evaluate the robustness of trackers on a range of test scenarios representing several real-world conditions. To compare single-target trackers' performance on trials, we present a single-score parameter-independent evaluation measure that quantifies tracking success and failure, and combines them for both summative and formative performance assessment. The protocol is validated on publicly available sequences with a diversity of targets and challenges using eight state-of-the-art single-target trackers. Through an extensive experimental analysis, the framework facilitates the selection of trackers for different operational conditions in real-world applications and for different target types.

Finally, to quantitatively compare the relative performances of evaluation measures we propose a methodology based on determining the probabilistic agreement between tracking result

decisions made by measures and those made by humans. We use tracking results on publicly available datasets with different target types and varying challenges, and collect the judgments of 90 skilled, semi-skilled and unskilled human subjects using a web-based performance assessment test. The analysis of the agreements allows us to highlight the variation in performance of the different measures and the most appropriate ones for the evaluation and comparison of trackers.

Contents

Published work

Journal papers

[J1] T. Nawaz, F. Poiesi and A. Cavallaro. Measures of effective video tracking. *IEEE Transactions on Image Processing*, Vol. 23, Issue 1, pp. 376-388, January 2014.

[J2] T. Nawaz and A. Cavallaro. A protocol for evaluating video trackers under real-world conditions. *IEEE Transactions on Image Processing*, Vol. 22, Issue 4, pp. 1354-1361, April 2013.

[J3] T. Nawaz and G. Slabaugh. A bottom-up approach for the analysis of haustral fold ridges in CTC-CAD. *Annals of the British Machine Vision Association*, vol. 2012, no. 8, pp. 1-15, 2012.

Conference papers

[C1] T. Nawaz, A. Cavallaro and B. Rinner. Trajectory clustering for motion pattern extraction in aerial videos. In *Proc. of IEEE International Conference on Image Processing*, Paris, 27-30 October 2014.

[C2] T. Nawaz, F. Poiesi and A. Cavallaro. Assessing tracking assessment measures. In *Proc. of IEEE International Conference on Image Processing*, Paris, 27-30 October 2014.

[C3] T. Nawaz and A. Cavallaro. PFT: a protocol for evaluating video trackers. In *Proc. of IEEE International Conference on Image Processing*, Brussels, 11-14 September 2011.

[C4] T. Nawaz and G. Slabaugh. Global analysis of haustral fold ridges for the reduction of false positives in CTC-CAD. In *Proc. of Medical Image Understanding and Analysis*, London, 14-15 July 2011.

NB: Please note that [J3, C4] are not a part of this thesis.

Electronic preprints are available at http://www.eecs.qmul.ac.uk/~andrea/publications.html.

Glossary of abbreviations

AER	Accuracy Error Rate	54
ANOVA	ANalysis Of VAriance	83
AVSS	Advanced Video and Signal-based Surveillance	27
BeyondSemiBoost	Beyond semisupervised boosting tracker	77
Boost	Online boosting tracker	76
CAST	Centre for Applied Science and Technology	44
CAVIAR	Context Aware Vision using Image-based Active Recognition	19
CBWH	Corrected Background-Weighted Histogram based mean-shift tracker	77
CDT	Correct Detected Track	39
CER	Cardinality Error Rate	54
CHIL	Computers in the Human Interaction Loop	43
CLEAR	Classification of Events, Activities and Relationships	19
CoTPS	Combined Tracking Performance Score	68
CPNI	Centre for the Protection of National Infrastructure	44
CRFBT	Conditional Random Field based tracker	56
CT	Compressive tracker	77
CTR	Correct Track Ratio	34
ETISEO	Evaluation du Traitement et de l'Interpretation de Sequences vidEO	19
FAR	False Alarm Rate	37
FAT	False Alarm Track	39
FP	False Positive	38
FN	False Negative	39
FragTrack	Fragments-based tracker	77

HAREM	Human Activity Recognition and Modelling	42
IDC	Identity Changes	38
IDS	Identity Switches	40
ID	Identity	5
i-LIDS	imagery Library for Intelligent Detection Systems	19
ITU	International Telecommunication Union	89
JPEG	Joint Photographic Experts Group	70
MCMCDA	Markov-Chain Monte-Carlo Data Association algorithm	58
MD	Mean Dice	34
MELT	Multiple Extended-target Lost-Track ratio	41
METE	Multiple Extended-target Tracking Error	41
MLT	Mean Length of ground-truth Tracks with id change(s)	62
MODA	Multiple Object Detection Accuracy	39
MOTA	Multiple Object Tracking Accuracy	39
MOTP	Multiple Object Tracking Precision	39
MS	Mean-Shift tracker	73
MT-TBD	Multi-target Track-Before-Detect	58
N-MODA	Normalised Multiple Object Detection Accuracy	39
NIDC	Normalised Identity Changes	41
OSPA	Optimal Sub-Pattern Assignment	37
OTE	Object Tracking Error	36
PAPTE	Point-based Assignment and Position-based Tracking Evaluation	35
PETS	Performance Evaluation of Tracking and Surveillance	19
PF	Particle filter-based tracker	77
RAPTE	Region-based Assignment and Position-based Tracking Evaluation	35
RASTE	Region-based Assignment and Size-based Tracking Evaluation	35
SemiBoost	Semi-supervised online boosting tracker	76
SPEVI	Surveillance Performance EValuation Initiative	45

TDF	Track Detection Failure	39
TDR	Track Detection Rate	37
TF	Track Fragmentation	38
TP	True positive	38
TRDR	Tracker Detection Rate	37
TSP	Tracking Success Probability	33
VOT	Visual Object Tracking challenge	19
VSCA	Computer Vision System Control Architectures	42

Glossary of symbols

$\mathcal{V}$	video sequence	21
K	number of frames in $\mathcal{V}$	21
f_k	video frame k of $\mathcal{V}$	73
$X_{k,j}$	estimated state of target j at frame k	21
$x_{k,j}$	estimated x-coordinate of the position of target j at frame k	21
$y_{k,j}$	estimated y-coordinate of the position of target j at frame k	21
$A_{k,j}$	estimated region information of target j at frame k	21
l_j	estimated ID of target j	21
$X'_{k,j}$	estimated state of target j at frame k excluding $A_{k,j}$	21
$\mathbf{X}_k$	set of estimated states of multiple targets at frame k	21
u_k	number of estimated targets at frame k	21
$\mathfrak{X}_j$	estimated track of target j	21
k^j_{ini}	initial frame number of $\mathfrak{X}_j$	21
k^j_{end}	final frame number of $\mathfrak{X}_j$	21
K_j	number of frames spanned by $\mathfrak{X}_j$	21
$\mathcal{X}$	set of estimated trajectories	21
U	number of estimated tracks $\mathcal{X}$	21
$\bar{X}_{k,i}$	ground-truth state of target i at frame k	22
$\bar{x}_{k,i}$	ground-truth x-coordinate of position of target i at frame k	22
$\bar{y}_{k,i}$	ground-truth y-coordinate of position of target i at frame k	22
$\bar{A}_{k,i}$	ground-truth region information of target i at frame k	22
$\bar{l}_i$	ground-truth ID of target i	22
$\bar{X}'_{k,i}$	ground-truth state of target i at frame k excluding $\bar{A}_{k,i}$	22
$\bar{\mathbf{X}}_k$	set of ground-truth states of multiple targets at frame k	22
v_k	number of ground-truth targets at frame k	22
$\bar{\mathfrak{X}}_i$	estimated track of target i	22

$\bar{k}^i_{ini}$	initial frame number of $\bar{\mathfrak{X}}_i$	22
$\bar{k}^i_{end}$	final frame number of $\bar{\mathfrak{X}}_i$	22
$\bar{K}_i$	number of frames spanned by $\bar{\mathfrak{X}}_i$	22
$\bar{\mathcal{X}}$	set of ground-truth trajectories	22
V	number of ground-truth tracks $\bar{\mathcal{X}}$	22
d_k	distance between X_k and $\bar{X}_k$	31
d_{avg}	average distance between $\mathfrak{X}$ and ground-truth trajectory $\bar{\mathfrak{X}}$	31
O_k	amount of overlap at frame k	32
D_k	dice score at frame k	32
TSP_k	tracking success probability at frame k	33
ν	parameter in TSP measure based on overlap threshold τ_1	33
$a(\bar{A}_k, A_k)$	overlap between $\bar{A}_k$ and A_k in TSP	33
$\hat{P}$	precision	33
$\hat{R}$	recall	33
τ_2	overlap threshold used $\hat{P}$ and $\hat{R}$	33
$\mathcal{F}$	F-score	33
AUC_λ	area under the lost track ratio curve	34
$\lambda(\tau)$	lost-track ratio curve corresponding to the varying τ	34
$\mathrm{OTE}_{\grave{t}}$	object tracking error between the estimated and ground-truth track pair $\grave{t}$	34
$\hat{K}_{\grave{t}}$	number of common frames in the associated track pair $\grave{t}$	36
$W_p(\bar{\mathbf{X}}_k, \mathbf{X}_k)$	Wasserstein's distance between $\mathbf{X}_k$ and $\bar{\mathbf{X}}_k$	36
$d(X'_{k,j}, \bar{X}'_{k,i})^p$	distance (p-norm) between $X'_{k,j}$ and $\bar{X}'_{k,i}$	36
$\mathcal{D}_{p,c}(\bar{\mathbf{X}}_k, \mathbf{X}_k)$	OSPA distance between $\bar{\mathbf{X}}_k$ and $\mathbf{X}_k$	37
Π_{u_k}	set of permutations taken from $\{1, 2, \ldots, u_k\}$ in OSPA	37
$\grave{D}_c(\bar{X}', X')$	cut-off distance in OSPA	37
c	cut-off parameter in OSPA	37
p	order parameter in OSPA	37
$\grave{D}(\bar{X}', X')$	base distance between $\bar{X}'$ and X' in OSPA	37
TRDR_k	tracker detection rate at frame k	37
FAR_k	false alarm rate at frame k	38

TDR_i	track detection rate corresponding to ground-truth track i	38
$\widehat{TP}_k$	true positive estimations at frame k	37
$\widehat{FP}_k$	false positive estimations at k	38
$\widehat{TP}_j$	true positive estimations in $\mathfrak{X}_j$	38
TF_i	track fragmentation measured for ground-truth track $\bar{\mathfrak{X}}_i$	38
IDC_i	ID changes measured with respect to ground-truth track $\bar{\mathfrak{X}}_i$	38
τ_o	overlap threshold used in MOTP	40
MODA_k	multiple object detection accuracy at frame k	40
$\mathcal{A}_k$	accuracy error	53
$\mathcal{C}_k$	cardinality error	53
METE_k	multiple extended-target tracking error at frame k	53
MELT_τ	multiple extended-target lost-track ratio value corresponding to τ	56
λ_i^τ	lost-track ratio at τ for the associated pair of ground-truth track i and estimated track(s)	55
N_i	number of frames in the ground-truth track i	55
N_i^τ	number of frames having overlap $O(\cdot) \leq \tau$	55
$\hat{S}_\tau$	number of sampled τ values	56
IDC_i^{max}	maximum number of ID changes corresponding to ground-truth track i	59
NIDC_i	normalised ID changes value for the ground-truth track i	59
V_{IDC}	number of ground-truth tracks having ID change(s)	59
$T_{\hat{j}}$	tracker $\hat{j}$	69
$P_{\hat{i}}$	trial $\hat{i}$	69
h_t	target t	69
I_t	tracker's initialisation for target t	69
$\mathcal{V}_t$	video sequence containing the target t	69
$I_{t,\hat{i}}$	tracker's initialisation for target t on trial $\hat{i}$	69
$\mathcal{V}_{t,\hat{i}}$	test sequence with target t generated on trial $\hat{i}$	69
$\mathfrak{X}_{t,\hat{i}}^{\hat{j}}$	trajectory obtained by running a tracker $T_{\hat{j}}$ with initialisation $I_{t,\hat{i}}$ and sequence $\mathcal{V}_{t,\hat{i}}$	69

n_1	number of initialisation perturbations generated on trial 1	70
n_2	number of initialisation perturbations generated on trial 2	70
n_3	number of initialisation perturbations generated on trial 3	70
n_4	number of generated test sequences on trial 4	70
n_5	number of generated test sequences on trial 5	70
$\hat{m}-1$	number of frames perodically dropped on trial 5	70
n_6	number of generated test sequences on trial 6	70
ΔL	change in illumination level on trial 6	70
n_7	number of generated test sequences on trial 7	70
ζ	compression quality parameter on trial 7	70
n_8	number of generated test sequences on trial 8	71
ρ	percentage reduction in the resolution of frame on trial 8	71
$\hat{N}^{\hat{\tau}}$	number of frames with overlap $O_k > 0$ and $O_k < \hat{\tau}$ with $\hat{\tau}$ being an overlap threshold	73
$\hat{N}$	number of frames with $O_k \neq 0$	73
$\hat{\lambda}^{\hat{\tau}}$	lost-track ratio computed using $\hat{F}^{\hat{\tau}}$ corresponding to $\hat{\tau}$	73
Ω	tracking accuracy computed in CoTPS	73
λ_0	tracking failure computed in CoTPS	74
N^0	number of frames with $O_k = 0$	74
β	weighting parameter in CoTPS	74
μ_C	mean CoTPS	77
d_C	difference between maximum and minimum CoTPS values of a tracker on a trial	77
α	significance level in the statistical significance test	83
$\mathfrak{X}_{\grave{i}}^1$	estimated trajectory of tracker 1, T_1, on video clip $\grave{i}$, $\mathcal{V}_{\grave{i}}$, with initialisation $I_{\grave{i}}$	86
$S_{\grave{i}\grave{j}}^1$	performance score computed using the measure $\grave{j}$ by evaluating $\mathfrak{X}_{\grave{i}}^1$ with respect to ground-truth trajectory $\bar{\mathfrak{X}}_{\grave{i}}$	86
$R_{\grave{i}\grave{j}}$	ranking of the measure $\grave{j}$ based on score of tracker 1, $S_{\grave{i}\grave{j}}^1$, and tracker 2, $S_{\grave{i}\grave{j}}^2$	86

$R_{\grave{i}\grave{l}}$	ranking of the human subject $\grave{l}$ based on $\mathfrak{X}_{\grave{i}}^{1}$ and $\mathfrak{X}_{\grave{i}}^{2}$ while using also the shown ground-truth samples, $\bar{\mathfrak{X}}_{\grave{i}}^{sam}$	86
$\overline{O}$	average of overlap, O_k, across the sequence	87
$\overline{\text{TSP}}$	average of TSP scores across the sequence	87
$CTR_{0.7}$	correct track ratio corresponding to mean dice of atleast 0.7	87
$\grave{N}_1$	number of skilled subjects used in assessing measures	89
$\grave{N}_2$	number of semi-skilled subjects used in assessing measures	89
$\grave{N}_3$	number of unskilled subjects used in assessing measures	89
χ^2	Friedman's test statistics	90
$\mathbf{E}^{\grave{i}}$	set containing events $(E_1^{\grave{i}}, E_2^{\grave{i}}, E_3^{\grave{i}})$ for a sample of subjects in a probability space for each $\mathcal{V}_{\grave{i}}$	91
$P(E_1^{\grave{i}})$	probability of occurrence of event $E_1^{\grave{i}}$	92
$n_{E_1^{\grave{i}}}$	number of times $E_1^{\grave{i}}$ occurs for each $\mathcal{V}_{\grave{i}}$	92
$P(B_{\grave{j}})$	probability of the agreement of $\grave{j}$th measure with a sample of human subjects	92
$P(B_{\grave{j}}^{\grave{i}} \vert E_1^{\grave{i}})$	probability of occurrence of $B_{\grave{j}}^{\grave{i}}$ given $E_1^{\grave{i}}$	93

Chapter 1

Introduction

1.1 Motivation

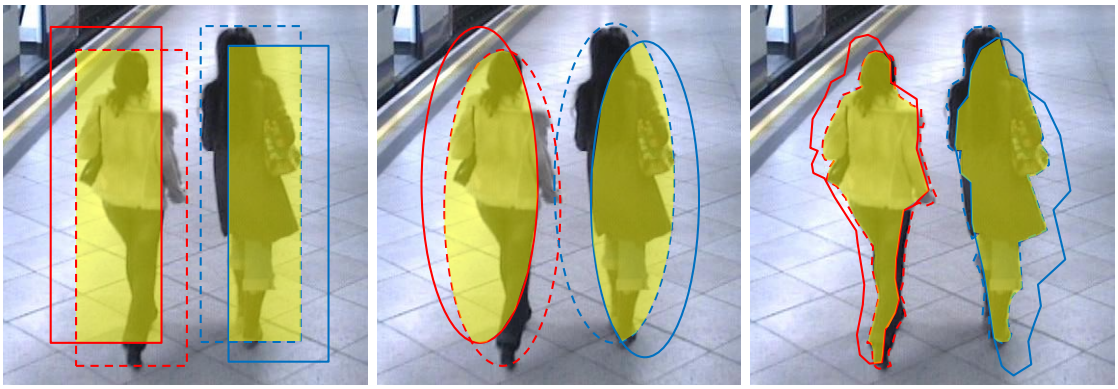
Video tracking is a widely-researched topic and is used in several applications such as activity analysis [108, 120], behaviour analysis [88, 119], abnormality detection [18, 17], event recognition [57, 50], people counting [37, 12], path detection [63, 64], motion pattern extraction [45, C1] and human-computer interaction [97, 16]. The choice of a video tracker for a specific application or task is still challenging because of the lack of an effective evaluation procedure that is capable of highlighting strengths and weaknesses of different trackers. In fact, unlike other areas of image processing and computer vision that consistently use commonly-accepted evaluation procedures such as disparity estimation [92], optical flow computation [5] and video coding [47], video tracking still lacks a standard way to evaluate and compare algorithmic performance. Although a number of efforts have been made towards performance evaluation of trackers in the form of evaluation campaigns (ETISEO[1], CLEAR[2], PETS[3], i-LIDS[4], CAVIAR[5], VOT Challenge[6]) and small-scale evaluation frameworks ([6, 10, 13, 23, 30, 52, 58, 60, 72, 76, 84, 89]), the performance of trackers is still tested using different evaluation criteria and varying datasets, thus hindering an effective evaluation and comparison.

The output of a video tracker is a trajectory or set of trajectories. A trajectory is defined as a

[1] http://www-sop.inria.fr/orion/ETISEO/index.htm. Accessed March 2014.
[2] http://www.clear-evaluation.org/. Accessed May 2011.
[3] http://www.cvg.rdg.ac.uk/slides/pets.html. Accessed March 2014.
[4] http://www.ilids.co.uk. Accessed March 2014.
[5] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. Accessed March 2014.
[6] http://www.votchallenge.net/index.html. Accessed March 2014.

sequence of target states estimated in the image plane over time using a tracker, where a tracker may aim to track only a single target, *single-target tracking* [2, 25, 118], or multiple targets in the scene, *multi-target tracking* [8, 81, 116]. The state may contain the information about target position or may also use the information about its occupied region in the image plane. Trajectory evaluation may involve the computation of the discrepancy over time between the estimated and the corresponding ground-truth states [6, 52, 86, 95] (Fig. 1.1). The discrepancy is computed using distance-based measures [67, 72, 86, 95] or overlap-based measures [10, 52, 89, 117]. *Distance-based measures* use the concept of distance minimisation between estimated and ground-truth states to evaluate performance. The distance-based evaluation might not consider the changes in the target size in the evaluation procedure [86, 72, 67] or the computed distance scores might not explicitly detect instances of tracking failure [95, 72, 67], where a tracking failure refers to the case of zero-overlap between estimated and corresponding ground-truth states. *Overlap-based measures* quantify the amount of common pixels (overlap) between the estimated and corresponding ground-truth states. The overlap-based evaluation generally takes into account target-size variations and can therefore detect instances of tracking failure [89, 52, 117]. Both distance-based and overlap-based measures may need presetting of parameters [52, 86]. For example, a cut-off parameter is used to define an upper bound [86]. Then the false positive estimations (i.e. incorrect estimations) and false negative estimations (i.e. missed estimations) are determined by comparing their spatial overlaps with a pre-defined threshold [52]. Moreover, some existing measures are numerically unbounded [52, 44] and not well defined for the worst tracking case.

It can be noticed that several performance evaluation measures have been introduced to measure the quality of video tracking results. These evaluation measures need, in turn, to be assessed in order to understand their relative performances. While efforts have been made to empirically assess measures in other research areas, including information retrieval [15], data clustering [53] and image compression [66], the direct quantitative assessment of measures in the area of video tracking is missing. A comparison of measures was performed indirectly by considering the performance of algorithms [7] and by studying the inter-measure correlation [65] without explicitly analysing the performance of the measures.



(a) (b) (c)

Figure 1.1: Tracking discrepancy is illustrated in the form of the overlap (yellow area) between estimated (solid line) and corresponding ground-truth (dotted line) states of targets. The target state includes the information about its position and occupied region in the image plane (extended-target representation), which can be defined as (a) bounding box, (b) bounding ellipse or (c) contour. Image is from the iLids Easy sequence.

1.2 Problem formulation

Let $X_{k,j}$ be the state of target j estimated by a tracker at frame k of the video sequence $\mathcal{V}$ and defined as

$$X_{k,j} = (x_{k,j}, y_{k,j}, A_{k,j}, l_j), \tag{1.1}$$

where $(x_{k,j}, y_{k,j})$ define the position of the target, $A_{k,j}$ is its region information on the image plane and l_j is the target ID, and $k = 1, \ldots, K$. K is the number of frames in the video sequence. When the estimated state of the target j at frame k contains only $(x_{k,j}, y_{k,j})$, it is denoted as $X'_{k,j}$. $\mathbf{X}_k$ is the set of estimated states of multiple targets:

$$\mathbf{X}_k = \{X_{k,1}, \ldots, X_{k,j}, \ldots, X_{k,u_k}\}, \tag{1.2}$$

where $u_k = |\mathbf{X}_k|$ is the number of estimated targets at frame k (i.e. the cardinality of $\mathbf{X}_k$). The trajectory $\mathfrak{X}_j$ of target j (also referred to as a *track* in the literature) is defined as a sequence of states over time:

$$\mathfrak{X}_j = \{X_{k,j}\}_{k=k^j_{ini}}^{k^j_{end}}, \tag{1.3}$$

where k^j_{ini} and k^j_{end} denote the initial and final frame numbers of $\mathfrak{X}_j$, respectively, and K_j is the number of frames spanned by $\mathfrak{X}_j$. $\mathcal{X}$ is the set containing all the estimated tracks in the sequence:

$$\mathcal{X} = \{\mathfrak{X}_j\}_{j=1}^{U}, \tag{1.4}$$

where U denotes the number of estimated tracks. Similarly, $\bar{X}_{k,i}$, $(\bar{x}_{k,i}, \bar{y}_{k,i}, \bar{A}_{k,i}, \bar{l}_i)$, $\bar{\mathbf{X}}_k$, v_k, $\bar{X}'_{k,i}$, $\bar{\mathfrak{X}}_i$, $\bar{k}^i_{ini}$, $\bar{k}^i_{end}$, $\bar{K}_i$, $\bar{\mathcal{X}}$ and V are the corresponding ground-truth notations for $X_{k,i}$, $(x_{k,i}, y_{k,i}, A_{k,i}, l_i)$, $\mathbf{X}_k$, u_k, $X'_{k,j}$, $\mathfrak{X}_j$, k^j_{ini}, k^j_{end}, K_j, $\mathcal{X}$ and U, respectively.

Single-target tracking evaluation involves simply quantifying the deviation of $\mathfrak{X}$ with respect to $\bar{\mathfrak{X}}$ into the evaluation score, $\mathcal{S}$, [60, 89]. Note that the subscripts j and i are removed from the notations for simplicity since $U = V = 1$ for the case of single-target tracking. Alternatively, multi-target tracking evaluation first involves determining the assignment between estimated and ground-truth states [86, 10, 52, 117, 13]. After the assignment, $\mathcal{X}$ is evaluated with respect to $\bar{\mathcal{X}}$ to obtain $\mathcal{S}$.

1.3 Contributions

Given a set of target trajectories estimated by performing tracking in a video sequence, we aim to devise criteria to quantitatively evaluate the estimated trajectories with respect to the ground-truth trajectories. To this end, we propose measures and procedures to provide an objective and holistic performance evaluation of trackers. Additionally, we propose a methodology to quantitatively assess the evaluation measures *per se*. Specifically, the following are the main contributions of the thesis:

1. Existing multi-target tracking evaluation measures do not evaluate target-size changes [86, 10, 44], rely on fixed thresholds [86, 52, 117], are not numerically bounded [86, 52], and do not account for cardinality error [10, 44]. We propose three threshold-independent and numerically-bounded measures for the performance evaluation of multi-target video trackers [J1][7]. The measures are overlap-based and account for the variations in target size over time. They provide frame-level evaluation by combining accuracy and cardinality errors, sequence-level accuracy evaluation based on the lost-track ratio information, and evaluation in terms of the ID changes normalised by the length of the track in which they occur. An extensive experimental validation is provided for the proposed measures by comparing them with the state-of-the-art measures and by evaluating four recent multi-target trackers on challenging real-world datasets. We made available online[8] the software implementing the measures to facilitate their use for the community. These measures would be useful for conducting a holistic evaluation of different aspects of multi-target tracking.

[7] The work in [J1] was jointly performed with F. Poiesi.

[8] http://www.eecs.qmul.ac.uk/~andrea/mtte.html. Accessed May 2014.

2. The real-world conditions, under which trackers operate in real applications, are not explicitly considered in the evaluation of trackers. These conditions refer to the distortions induced to the input of a tracker, which may affect its performance, such as initialisation perturbations, noisy and compressed video data, frame dropping, and illumination changes. We propose a protocol with a comprehensive set of trials that test the *robustness* of trackers on a range of test scenarios representing these real-world operational conditions [J2, C3]. To quantify the tracking performance on the trials, we propose a threshold-independent overlap-based criterion that summarises single-target tracking performance based on a new evaluation measure [J2], which takes into account target size variations. The proposed measure quantifies how *accurately* and how *long* a target is tracked across a sequence. To the best of our knowledge, this is the first initiative that enables evaluating the robustness of the performance of tracking algorithms under several real-world conditions. We performed an extensive validation of the protocol by evaluating and highlighting the strengths and weaknesses of eight state-of-the-art trackers using a set of sequences (and their variations) with a diversity of targets and challenges on more than 187000 frames ($\approx$ 3.25 hours of video data). The resulting performance evaluation tool is made available online as an open source software[9]. The website enables the researchers to compare and share the results of their trackers under several test scenarios. This work [J2, C3] has been helpful to other works [19, 79, 65, 22, 28, 35] and VOT Challenge 2013 [54, 55], and we expect it to be more useful for researchers in the future.

3. While several tracking evaluation measures exist, the question of how to quantitatively assess the measures *per se* is not addressed. To address this limitation we propose a methodology for the quantitative assessment of discrepancy-based evaluation measures with respect to human judgement [C2]. The comparison and analysis are based on determining the probabilistic agreement between the decisions made by measures and those made by humans on tracking results. Other works exist that utilised human judgements for different tasks including the efficient generation of large-scale video annotations [104, 105] and assessment of the estimated quality of the performed actions in videos [82]. We show the usefulness of the proposed methodology by assessing seven state-of-the-art measures on tracking results generated on ten publicly-available datasets with three target types (head, full body,

[9]http://www.eecs.qmul.ac.uk/~andrea/pft2. Accessed May 2014.

vehicle). This study helped in determining strengths and weaknesses of different measures, which could pave the way to improve them in the future. Moreover, the idea of the proposed methodology of assessing tracking evaluation measures with respect to human judgements could also be applied to the similar problem in other research areas.

1.4 Organisation of the thesis

This thesis is organised as follows:

Chapter 1: The introduction and motivation for the thesis are described in Sec. 1.1, followed by the problem formulation in Sec. 1.2. The contributions of the thesis are discussed in Sec. 1.3.

Chapter 2: The introduction to the chapter is provided in Sec. 2.1. A review of the state-of-the-art on non-ground-truth-based trajectory evaluation is presented in Sec. 2.2 and on ground-truth-based trajectory evaluation in Sec. 2.3. This is followed by an overview of the key campaigns (Sec. 2.4) and datasets (Sec. 2.5) for trackers' evaluation. A discussion on the limitations of the existing related work is presented in Sec. 2.6.

Chapter 3: The introduction to the chapter is provided in Sec. 3.1. The description of the three proposed multi-target tracking evaluation measures that quantify different aspects of tracking is given in Sec. 3.2, Sec. 3.3 and Sec. 3.4. In Sec. 3.5 the experimentation is provided that describes the datasets and trackers used, and discusses the advantages of the proposed measures and their use for the performance comparison of multi-target trackers. The chapter is summarised in Sec. 3.6.

Chapter 4: Sec. 4.1 presents an introduction to the chapter. This is followed by a problem definition (Sec. 4.2), description of the trials of the proposed protocol used to test trackers' robustness (Sec. 4.3) and description of the proposed single-target evaluation measure used to quantify trackers' results on trials (Sec. 4.4). Sec. 4.5 provides the experimental analysis and validation including the description of datasets and trackers used, trial-wise and target-wise performance comparison of single-target trackers, and a discussion on the overall performance of trackers. The summary of the chapter is given in Sec. 4.6.

Chapter 5: Sec. 5.1 provides the introduction to the chapter followed by the problem definition in Sec. 5.2. Sec. 5.3 describes the state-of-the-art measures to be assessed. The description of the subjective evaluation with respect to which the measures are to be assessed is provided in

Sec. 5.4. The proposed methodology to quantitatively assess the measures and the experimental results and analysis are explained in Sec. 5.5. The chapter is summarised in Sec. 5.6.

Chapter 6: The chapter presents a summary of the achievements of the thesis (Sec. 6.1) and future directions of work (Sec. 6.2).

Chapter 2

Related work

2.1 Introduction

A trajectory describes the evolution of the state of a target over time estimated by a tracker on the image plane. The state may include the target position, *point-target representation* (e.g. in feature-point tracking [102, 14]) (Fig. 2.1(a)), or may also use the information about the occupied region of the target in the image plane, *extended-target representation* (e.g. in face or person tracking [107, 100, 11, 116]). In the case of extended-target representation, the region information may be represented as a bounding box [11], a bounding ellipse [116] or a bounding contour [101] (Fig. 2.1(b-d)). The criteria for the evaluation of trajectories can be categorised into non-ground-truth-based and ground-truth-based. While the former category of methods evaluates tracker's performance without using ground-truth information [24, 20, 113, 91], the latter category of methods evaluates trackers' performance by quantifying the discrepancy between the estimated and ground-truth states over time [52, 60, 86, 95].

This chapter presents a review of the related work on trajectory evaluation. First we discuss the existing non-ground-truth-based trajectory evaluation methods (Sec. 2.2). This is followed by a review of the state of the art on ground-truth-based trajectory evaluation in Sec. 2.3. Additionally, we also review the existing campaigns and projects (Sec. 2.4), and discuss the important datasets (Sec. 2.5) introduced for tracking performance assessment.

(a) (b) (c) (d)

Figure 2.1: Representation of the state of a target (vehicle) as a (a) point, (b) bounding box, (c) bounding ellipse and (d) bounding contour. Cropped image is from the AVSS Challenge 2007 dataset.

2.2 Non-ground-truth-based trajectory evaluation

Non-ground-truth-based trajectory evaluation involves the use of information about the output or internal stages of a tracker at the current frame only, *single-frame-based evaluation criteria* [24, 31, 93], between two (consecutive or non-consecutive) frames, *two-frame-based evaluation criteria* [4, 112, 21], or across a sequence of frames, *multi-frame-based evaluation criteria* [93, 113, 91], to assess tracking performance. We discuss below the single-frame-based, two-frame-based and multi-frame-based evaluation criteria.

2.2.1 Single-frame-based evaluation criteria

Collins *et al.* [24] used the difference between the colour features (histograms of red, green and blue channels) of the foreground and background to determine the quality of estimated tracking result at a frame. Han *et al.* [42] used the same idea of foreground-background separability to estimate the track quality by using also the gradient orientation histogram information with the colour feature information. Other methods exist that evaluate tracking quality based on the contrast in the colour information [31] or motion vectors [31, 32] between the foreground and background computed along the contour of the tracked target. The aforementioned approaches work under the assumption of separability between foreground and background that may not be valid in all scene types. Alternatively, Schubert *et al.* [93] proposed an approach to quantify performance by comparing the estimated state of a target with the corresponding measurement in the measurement space for a Bayesian filtering tracking framework. This method of evaluating tracking may have inaccuracies in cluttered scenes where the measurements can be unreliable. A limitation in the above criteria is that they may not identify the tracker recovery after a failure and the re-occurrence of a failure after the recovery.

2.2.2 Two-frame-based evaluation criteria

Several methods exist that evaluate tracking quality by enforcing the smoothness in the motion, appearance or tracking uncertainty between consecutive frames. Specifically, these methods provide tracking evaluation by quantifying the inter-frame variation in the target speed [27, 114, 21, 29], direction [114, 21], shape [21], area [27, 21], colour [31, 69, 21] and texture [27] information, or by checking the inter-frame consistency in the spatial uncertainty of a tracker estimated as a covariance of the particles in a particle filtering framework [91]. These criteria consider an abrupt change in the evaluation score as the occurrence of a tracking failure. The use of the inter-frame evaluation score may not be able to detect the tracker recovery after a failure and the re-occurrence of a failure. Additionally, the use of tracker spatial uncertainty in [91] as an evaluation criterion is tracker dependent, which is addressed in [90] to make it applicable also for other trackers. Other methods exist that instead apply the consistency at the current frame with respect to the initialising frame (of the target) and can therefore identify the tracker recovery and the re-occurrence of a tracking failure. These criteria quantify tracking quality based on the similarity in the shape [114], area [114], colour [114], posterior density estimate [112] (in a particle filtering framework) and tracker uncertainty (based on observation likelihood in a particle filtering framework) [4] between the current frame and the initialising frame. The criteria in [114] make a hidden assumption of constancy in target shape, area and colour across the whole sequence (due to the similarity check with respect to the initialising frame), which may not be valid in the scenes with target scale and pose changes. Additionally, the criteria in [112, 4] are applicable for particle filtering-based trackers.

2.2.3 Multi-frame-based evaluation criteria

Tracking quality is determined based on scene-specific criteria that require prior semantic scene information [41, 80, 21]. Hall [41] and Piciarelli *et al.* [80] computed the goodness of the estimated trajectory based on its fitness in a reference scene model learned *a priori*. The non-termination of a trajectory at the scene exit zone (known *a priori*) is employed as a performance indicator to determine the occurrence of a tracking failure [21]. Alternatively, the information about the temporal length of trajectories is used as a part of tracking evaluation procedure [21, 93], where a shorter temporal length alludes to the potential occurrence of a tracking failure. An important cause of the shorter temporal length may be the target occlusion and a knowledge

Table 2.1: Summary of the existing non-ground-truth-based trajectory evaluation approaches, which include single-frame-based (SFB), two-frame-based (TFB) and multi-frame-based (MFB) criteria.

Ref.	Category	Evaluation procedure	Information used
[24]	SFB	Foreground-background separability	Colour
[42]	SFB	Foreground-background separability	Colour, gradient orientation
[32]	SFB	Foreground-background separability	Colour, motion vectors
[93]	SFB	Estimate-Measurement match, trajectory length check	Sensor measurements, trajectory information
[31]	SFB, TFB	Foreground-background separability, inter-frame consistency	Motion vectors, colour
[27]	TFB	Inter-frame consistency	Speed, area, texture
[29]	TFB	Inter-frame consistency	Speed
[69]	TFB	Inter-frame consistency	Colour
[21]	TFB	Inter-frame consistency, scene model fitting, trajectory length check	Speed, direction, shape, area, colour, trajectory information
[114]	TFB	Inter-frame consistency, Current frame-Initial frame consistency	Speed, direction, shape, colour
[112]	TFB	Current frame-Initial frame consistency	Posterior density estimate
[4]	TFB	Current frame-Initial frame consistency	Spatial uncertainty
[91, 90]	TFB, MFB	Inter-frame consistency, time-reversibility constraint	Spatial uncertainty, state estimate
[41]	MFB	Scene model fitting	Trajectory information
[80]	MFB	Scene model fitting	Trajectory information
[113]	MFB	Time-reversibility constraint	State estimate

of occlusion regions in the scene [46] could therefore improve tracking performance. Other criteria exist that evaluate tracking performance over time based on the time-reversibility constraint [113, 91]. At each frame, the process involves performing reverse tracking towards the target starting frame by initialising the tracker with the current state estimate. The similarity between the estimated state at the starting frame obtained as a result of reverse tracking and the prior initial state (at the starting frame) provides the tracking performance. The use of time reversibility can identify tracker recovery and re-occurrence of a tracking failure; however, this criterion would be desirable for the case of short sequences due to issues of error accumulation over time and computational expensiveness [91]. Tab. 2.1 presents a summary of the non-ground-truth-based trajectory evaluation criteria.

Non-ground-truth-based evaluation criteria are particularly useful when the generation of the ground truth information is infeasible or cumbersome, or for improving online the performance of trackers [9, 51]. Non-ground-truth-based evaluation is however out of the scope of this thesis. This thesis focuses on the ground-truth-based evaluation, which is desirable for obtaining a confident and repeatable performance assessment and comparison of trackers due to the availability of a known performance benchmark [38]. Next, we provide a detailed review of the related work on ground-truth-based trajectory evaluation.

2.3 Ground-truth-based trajectory evaluation

The evaluation of trajectories may be performed by comparing them with respect to the ground-truth trajectories [60, 52]. For the case of single-target tracking, the trajectory evaluation involves quantifying the closeness between the estimated and ground-truth states [72, 60]. For the case

(a) (b) (c)

Figure 2.2: Illustration of the key aspects of multi-target tracking evaluation including accuracy, cardinality error and ID changes. Ground-truth states of targets are shown as dotted bounding boxes and estimated states are shown as solid bounding boxes. Different colours show the unique IDs of the targets. (a) Accuracy is illustrated in the form of the extent of overlap (shown as yellow shading) between the corresponding pairs of estimated and ground-truth bounding boxes; (b) cardinality error occurs because the number of estimated targets is not equal to the number of ground-truth targets; (c) Three ID changes occurs since all targets have changed their IDs (one target gets reinitialised with a new ID and other two targets have swapped their IDs). Image is taken from the iLids Easy dataset.

of multi-target tracking evaluation, after solving the assignment between $\mathcal{X}$ and $\bar{\mathcal{X}}$ [86, 10, 52] the three key aspects to be considered are accuracy, cardinality and number of ID changes. The *accuracy* evaluates the extent of agreement between estimated and ground-truth states [1], which can be quantified in terms of a distance score [86], overlap score [10, 52], or true positive (correct), false positive (incorrect) and false negative (missed) estimations [62] (Fig. 2.2(a)). The *cardinality error* is defined as the error in computing the number of targets (Fig. 2.2(b)). *ID changes* refer to the wrong associations between estimated and ground-truth targets (Fig. 2.2(c)). Moreover, single-target and multi-target tracking evaluation operates at frame level [60, 86] or at sequence level [89, 52]. The sequence-level multi-target tracking evaluation can be performed by considering either individual tracks [10] or all the tracks [52]. We analyse the existing measures in terms of the criterion used (distance, amount of overlap), evaluation aspect (accuracy, cardinality error, ID changes), use(no use) of fixed thresholds, and whether evaluation is performed at frame level or sequence level.

In the remaining part of this section we shall review the state-of-the-art ground-truth-based trajectory evaluation measures. First we shall discuss the evaluation measures for single-target tracking where the assignment problem is not required to be solved. These measures use the distance, distance-based, (Sec. 2.3.1) or the amount of overlap, overlap-based, (Sec. 2.3.2)

between the estimated and ground-truth results for assessing the performance. Then, we shall discuss multi-target tracking evaluation where we explain the assignment problem (Sec. 2.3.3) followed by the description of multi-target tracking evaluation measures (Sec. 2.3.4, Sec. 2.3.5, Sec. 2.3.6).

2.3.1 Distance-based measures for single-target tracking evaluation

A simple way to evaluate tracking is to compute the target centroid position error between X'_k and $\bar{X}'^1_k$, which is the distance in pixels between the position of estimated and ground-truth states $((x_k, y_k), (\bar{x}_k, \bar{y}_k))$ yielding object positional accuracy [72, 96, 67]:

$$d_k = \sqrt{(x_k - \bar{x}_k)^2 + (y_k - \bar{y}_k)^2}. \tag{2.1}$$

Alternatively, tracking is evaluated by computing the distance (d_k) between X_k and $\bar{X}_k$ when the state parameters associated with the region information ($A_k, \bar{A}_k$) are also taken into acccount [62]. Tracking performance is also assessed based on averaging the computed distance values across the estimated trajectory ($\mathfrak{X}$) with respect to ground-truth trajectory ($\bar{\mathfrak{X}}$) [71, 43, 62].

$$d_{avg} = \frac{1}{\hat{K}} \sum_{k=k_{ini}}^{k_{end}} d_k, \tag{2.2}$$

where $\hat{K}$ is the number of frames where the target exists. Distance-based measures may not evaluate the changes in the target size [72, 67] or may not effectively determine tracking failure [95, 72, 67] (Fig. 2.3). The overlap-based evaluation generally includes the variations in the target size in the evaluation procedure and an instance of no overlap declares a tracking failure.

2.3.2 Overlap-based measures for single-target tracking evaluation

Overlap-based measures quantify in different forms the overlay between estimated region, A_k, and ground-truth region, $\bar{A}_k$, over a sequence [62, 73] or simply define them to be *coincident* when the centroid of any of the two lies inside the area of the other [10] (Fig. 2.4). In the former case, the evaluation is defined at pixel level at frame k. In the latter case, the evaluation involves classifying results at target-level or trajectory-level as true positives, $\widehat{TP}$, false positives, $\widehat{FP}$, true negatives, $\widehat{TN}$, and false negatives, $\widehat{FN}$ [10, 13].

[1]The subscripts i and j, which refer to the target number, are not needed since a single target is under consideration.

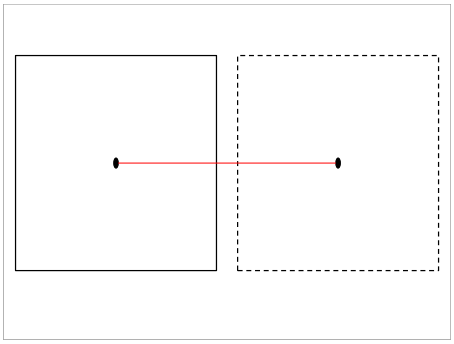


Figure 2.3: Illustration of occurrence of tracking failure due to the zero-overlap between the estimated bounding box (shown in solid lines) and ground-truth bounding box (shown in dotted lines). The distance (shown in red colour) between centroids of two bounding boxes is $\hat{a} > 0$. This value does not declare this case as a tracking failure *per se*; however, the amount of overlap between bounding boxes does.

The amount of *overlap* between A_k and $\bar{A}_k$ in a frame k can be given, when a target is present, by O_k [62]:

$$O_k = \frac{|\bar{A}_k \cap A_k|}{|\bar{A}_k \cup A_k|}, \tag{2.3}$$

where $O_k \in [0,1]$ and $|\cdot|$ represents the cardinality of a set. Alternatively, the *Dice* score, D_k, [73] gives more value to the correctly classified pixels and is computed as:

$$D_k = \frac{2|\bar{A}_k \cap A_k|}{|\bar{A}_k| + |A_k|}, \tag{2.4}$$

where $0 \leq D_k \leq 1$.

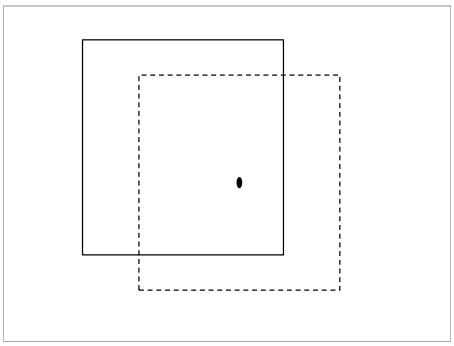


Figure 2.4: This figure shows coincidence between estimated bounding box (shown in solid lines) and ground-truth bounding box (shown in dotted lines) because the centroid position of ground-truth bounding box lies within the area of the estimated bounding box.

The Tracking Success Probability, TSP_k, is defined as [60]:

$$\text{TSP}_k = \frac{\exp(\nu \cdot a(\bar{A}_k, A_k))}{1+\exp(\nu \cdot a(\bar{A}_k, A_k))} \in [0,1], \tag{2.5}$$

where $a(\bar{A}_k, A_k)$ quantifies the overlap between $\bar{A}_k$ and A_k, and $\nu > 0$ is fixed *a priori* based on an application-specific overlap threshold, τ_1, which defines tracking as successful when $a(\bar{A}_k, A_k) \geq \tau_1$. When $\bar{A}_k = (\check{l}_g^k, \check{r}_g^k, \check{t}_g^k, \check{b}_g^k)$ and $A_k = (\check{l}_r^k, \check{r}_r^k, \check{t}_r^k, \check{b}_r^k)$ are defined by the horizontal or vertical coordinates of the left, right, top and bottom boundaries of the ground truth and the estimated bounding boxes, respectively, $a(\bar{A}_k, A_k)$ is given as follows [59]:

$$a(\bar{A}_k, A_k) = \check{S}_{rg}^k \cdot \left| \frac{min(\hat{\mathcal{H}}_k) \cdot min(\hat{\mathcal{V}}_k)}{max(\hat{\mathcal{H}}_k) \cdot max(\hat{\mathcal{V}}_k)} \right|, \tag{2.6}$$

where $\hat{\mathcal{H}}_k = \{\check{r}_r^k - \check{l}_g^k, \check{r}_g^k - \check{l}_r^k, \check{r}_g^k - \check{l}_g^k, \check{r}_r^k - \check{l}_r^k\}$, $\hat{\mathcal{V}}_k = \{\check{b}_r^k - \check{t}_g^k, \check{b}_g^k - \check{t}_r^k, \check{b}_g^k - \check{t}_g^k, \check{b}_r^k - \check{t}_r^k\}$, and $\check{S}_{rg}^k$ is an indicator function such that $\check{S}_{rg}^k = -1$, if A_k and $\bar{A}_k$ do not overlap, and $\check{S}_{rg}^k = 1$, otherwise.

The measures reviewed above offer information about how objects are tracked within a single frame (frame-level evaluation); however, since trackers are temporal filters, it is also relevant to evaluate their performance at sequence level across multiple frames. We discuss below the measures, which provides sequence-level evaluation [62, C3, 89].

Precision, $\hat{P}$, quantifies tracking performance by penalising the number of true positives (correct estimations), $|\widehat{TP}|$, with respect to the number of false positives (incorrect estimations), $|\widehat{FP}|$ [62]:

$$\hat{P} = \frac{|\widehat{TP}|}{|\widehat{TP}| + |\widehat{FP}|}, \tag{2.7}$$

where an estimation is a true positive if $O_k \geq \tau_2$ and a false positive if $O_k < \tau_2$. τ_2 is a pre-set threshold. Likewise, Recall, $\hat{R}$, penalises the $|\widehat{TP}|$ with respect to the number of false negatives (missed estimations), $|\widehat{FN}|$ [62]:

$$\hat{R} = \frac{|\widehat{TP}|}{|\widehat{TP}| + |\widehat{FN}|}. \tag{2.8}$$

The F-score is computed using Precision and Recall scores as follows [62]:

$$\mathcal{F} = 2\frac{\hat{P} \cdot \hat{R}}{\hat{P} + \hat{R}}. \tag{2.9}$$

Unlike $\hat{P}, \hat{R}, \mathcal{F}$, the evaluation criteria in [C3, 89] provide sequence-level tracking performance without relying on any preset threshold parameters. The measure [C3] uses the overlap

(a)	(b)	(c)	(d)	(e)
$O_k = 0.555$	$O_k = 0.559$	$O_k = 0.559$	$O_k = 0.555$	$O_k = 0.551$
$D_k = 0.714$	$D_k = 0.717$	$D_k = 0.717$	$D_k = 0.714$	$D_k = 0.711$
$\text{TSP}_k = 0.998$	$\text{TSP}_k = 0.998$	$\text{TSP}_k = 0.998$	$\text{TSP}_k = 0.998$	$\text{TSP}_k = 0.998$

Figure 2.5: Evaluation scores of single-target tracking measures for a sequence of five images (a) to (e). Estimated result: blue; ground truth: green. D_k is higher than O_k since the numerator in Eq 2.4 is multiplied by 2 thus giving more weighting to the number of common pixels between estimated result and ground truth. TSP_k is the same for all samples (a) to (e) due to its parameter dependency. Moreover, for the tracking example, $\hat{P} = 1$, $\hat{R} = 1$, $\mathcal{F} = 1$, $AUC_\lambda = 0.445$. For computing $\hat{P}$ and $\hat{R}$, we set $\tau_2 = 0.5$ [8]. Unlike $\hat{P}$, $\hat{R}$ and $\mathcal{F}$, AUC_λ can identify the discrepancy between the estimations and the ground truth.

information, O_k (Eq. 2.3), to evaluate the performance. Tracking performance is quantified into a single score by computing the *area under the lost track ratio curve* (AUC_λ) as follows:

$$AUC_\lambda = \Delta\tau \sum_{\tau=0}^{1} \lambda(\tau), \tag{2.10}$$

where $AUC_\lambda \in [0,1]$ and the lower AUC_λ, the better the tracking result. λ is the ratio of the number of frames with a lost track and the total frames in the estimated trajectory, where a track is considered lost if $O_k \leq \tau$ with $\tau \in [0,1]$ being the threshold value. Because the appropriate value of τ can be different for different tracking applications, the variation of λ is considered for a full range of τ values, from $\tau = 0$ to $\tau = 1$ with an increment of $\Delta\tau$, and the parameterized values of the lost-track ratio are referred to as $\lambda(\tau)$. The tracking evaluation methodology presented in [89] uses the *Dice* score D_k. The evaluation procedure involves computing *Correct Track Ratio*, CTR, and mean dice, MD. CTR is the percentage of frames where the D_k is greater than a threshold and MD is the average of D_k scores that are greater than this threshold. CTR and MD are computed for a full range of threshold values and the resulting plot (*MD vs. CTR*) is used to analyse the tracking performance. The plot needs to be inspected by an operator in order to determine the tracking performance. See Fig. 2.5 for numerical examples and Tab. 2.2 for a summary of the single-target evaluation measures.

Table 2.2: Summary of single-target tracking evaluation measures. The comparison is based on whether the evaluation measure is threshold independent, the target-size changes considered in the evaluation, the tracker's robustness is evaluated partially (P) or thoroughly (T) in the framework, the type of measures i.e. distance-based or overlap-based, and the evaluation is at frame level or sequence level.

Ref.	Threshold independence	Size-change evaluation	Robustness evaluation	Type	Evaluation type
[96]	✓	✓		Distance-based	Frame level
[72]	✓			Distance-based	Frame level
[67]	✓			Distance-based	Frame level
[62]		✓		Distance-based, Overlap-based	Frame level, Sequence level
[71]	✓			Distance-based	Sequence level
[43]	✓			Distance-based	Sequence level
[73]	✓	✓		Overlap-based	Frame level
[10]	✓			Overlap-based	Frame level
[60]		✓	P	Overlap-based	Frame level
[89]	✓	✓		Overlap-based	Sequence level
[C3]	✓	✓	P	Overlap-based	Sequence level
[J2]	✓	✓	T	Overlap-based	Sequence level

2.3.3 Assignment problem for multi-target tracking evaluation

The trajectory evaluation for multi-target tracking requires first solving the assignment problem, which associates the estimated tracks to the ground-truth tracks. The assignment problem may be solved using only the position information (point-based assignment) that involves minimising the distance between the estimated and ground-truth tracks [86, 10], or using also the region information (region-based assignment) that considers the overlap [52, 117] or coincidence [13] between the estimated and ground-truth regions. Moreover, the assignment may be determined at frame level [8] or at sequence level [86]. The solution to the assignment problem is followed by the quantification of tracking performance using measure(s).

We can identify three categories of multi-target tracking evaluation measures, which are: Point-based Assignment and Position-based Tracking Evaluation (PAPTE) measures, Region-based Assignment and Position-based Tracking Evaluation (RAPTE) measures, and Region-based Assignment and Size-based Tracking Evaluation (RASTE) measures. *PAPTE measures* use target position information both for solving the assignment (point-based assignment) and for providing the evaluation (position-based assignment) between the estimated and ground-truth tracks without taking into account size changes over time (Fig. 2.6). *RAPTE measures* also use target region information to solve the assignment (region-based assignment) and offer a position-based evaluation of tracking. *RASTE measures* employ the region-based assignment and evaluate by also considering the temporal size variations (size-based evaluation). Next, we discuss the PAPTE, RAPTE and RASTE measures.

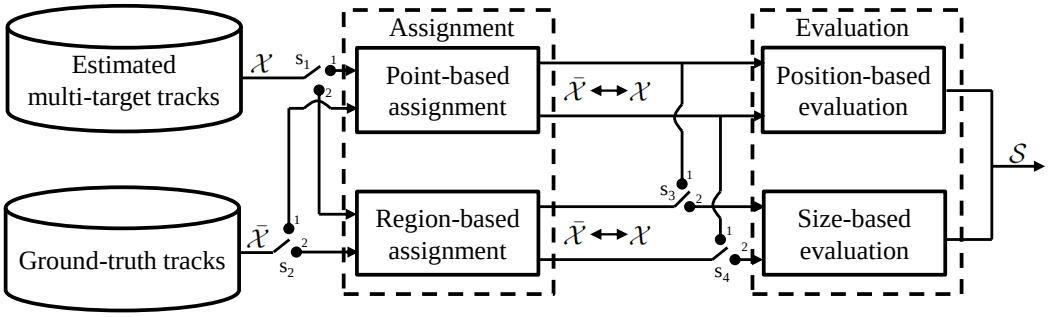


Figure 2.6: General procedure for the computation of the evaluation score $\mathcal{S}$ for multi-target tracking. Three different modalities are possible: using a point-based solution for the assignment problem and for evaluation ($s_1 = s_2 = 1, s_3 = s_4 = 2$); using a region-based solution for the assignment problem and a point-based solution for evaluation ($s_1 = s_2 = 2, s_3 = s_4 = 1$); using a region-based solution for the assignment problem and information about target position and size for evaluation ($s_1 = s_2 = 2, s_3 = s_4 = 2$).

2.3.4 Point-based assignment and position-based evaluation measures for multi-target tracking

PAPTE measures are Object Tracking Error [10], Wasserstein's distance-based metric [44], Optimal Sub-Pattern Assignment metric [86, 94], Tracker Detection Rate [10], False Alarm Rate [10], Track Detection Rate [10] and Track Fragmentation [10]. These measures are explained below.

Object Tracking Error (OTE) computes the tracking accuracy as the mean distance between the position of the estimated and ground-truth track pair. An estimated track is associated with ground-truth track based on the minimisation of the mean distance between their common frames [95]. Therefore, $\text{OTE}_{\hat{\imath}}$ is calculated for the associated pair $\grave{\imath}$ as follows:

$$\text{OTE}_{\hat{\imath}} = \frac{1}{\hat{K}_{\hat{\imath}}} \sum_{k=\hat{k}^{\grave{\imath}}_{ini}}^{\hat{k}^{\grave{\imath}}_{end}} \sqrt{(\bar{x}_{k,\grave{\imath}} - x_{k,\grave{\imath}})^2 + (\bar{y}_{k,\grave{\imath}} - y_{k,\grave{\imath}})^2}, \tag{2.11}$$

where $\hat{K}_{\hat{\imath}}$ denotes the number of common frames in the associated track pair $\grave{\imath}$, and equals $\hat{k}^{\grave{\imath}}_{end} - \hat{k}^{\grave{\imath}}_{ini}$.

The Wasserstein's distance computes the tracking accuracy between $\mathbf{X}_k$ and $\bar{\mathbf{X}}_k$ as follows:

$$W_p(\bar{\mathbf{X}}_k, \mathbf{X}_k) = \min_{\bar{\mathbf{C}}} \left(\sum_{j=1}^{u_k} \sum_{i=1}^{v_k} \bar{C}^k_{j,i} d(X'_{k,j}, \bar{X}'_{k,i})^p \right)^{1/p}, \tag{2.12}$$

where u_k and v_k are the number of estimated and ground-truth targets at frame k, respectively, $d(\cdot)^p$ defines the *p-norm* ($p \in [1, \infty)$), and the transportation matrix, $\bar{\mathbf{C}}$ contains the association costs between all possible estimated and ground-truth track pairs. The Hungarian or Munkres

algorithm [56, 70] is used to find the associations with the minimum overall cost.

Unlike OTE and $W_p(\cdot)$, the Optimal Sub-Pattern Assignment (OSPA) metric evaluates tracking by also including the cardinality error:

$$\mathcal{D}_{p,c}(\bar{\mathbf{X}}_k, \mathbf{X}_k) = \left[\frac{1}{max(u_k, v_k)} \left(\min_{\pi \in \Pi_{u_k}} \sum_{i=1}^{v_k} \left(\grave{D}_c(\bar{X}'_{k,i}, X'_{k,\pi(i)}) \right)^p + |u_k - v_k| \cdot c^p \right) \right]^{1/p}. \quad (2.13)$$

Π_{u_k} denotes the set of permutations such that the length of each permutation is v_k whose elements are taken from $\{1, 2, \ldots, u_k\}$; c is the cut-off parameter and defines the upper bound; p is the order parameter ($p \in [1, \infty)$) of OSPA metric, which penalises the estimated states that are far away from any of the ground-truth states; $\grave{D}_c(\bar{X}', X')$ is the cut-off distance between estimated and ground-truth states: $\grave{D}_c(\bar{X}', X') = \min(c, \grave{D}(\bar{X}', X'))$. The base distance between the two states, $\grave{D}(\bar{X}', X')$, combines localisation and labeling errors [86]:

$$\grave{D}(\bar{X}', X') = \left(||\bar{X}' - X'||_{\breve{p}} + \alpha^{\breve{p}} \delta[\bar{l}, l] \right)^{1/\breve{p}}, \quad (2.14)$$

where $\delta[\bar{l}, l] = 0$ when $\bar{l} = l$ and $\delta[\bar{l}, l] = 1$ when $\bar{l} \neq l$, and the parameter $\alpha \in [0, c]$ penalises the labelling error if the frame-level assignment established using the minimisation in Eq. 2.13 differs from the global assignment of tracks, which is determined *a priori* by minimising the average distance between estimated and ground-truth tracks [86, 36]. The parameter $\breve{p}$ defines the order of the base distance and typically $\breve{p} = p = 1$ [86].

Tracker Detection Rate (TRDR), False Alarm Rate (FAR) and Track Detection Rate (TDR) quantify the tracking accuracy while considering the information about true positives and false positives. While these measures take into account the target-size information (in determining true positives and false positives using the coincidence criterion), they are classified as PAPTE measures because they do not incorporate temporal target-size changes in the evaluation. The association between estimated and ground-truth tracks for TRDR, FAR and TDR is established as for OTE. TRDR provides the tracking performance at frame k as:

$$\text{TRDR}_k = \frac{|\widehat{TP}_k|}{v_k}, \quad (2.15)$$

where $|\widehat{TP}_k|$ is the number of true positive (correctly-tracked) targets and v_k is the number of ground-truth targets. With a target being represented as a bounding box, the true positive es-

timation refers to the coincidence of the centroid of the ground-truth bounding box with the estimated bounding box. An estimation is a false positive, $\widehat{FP}_k$, if the centroid of none of the ground-truth bounding boxes coincides(lies) with(in) an estimated bounding box. FAR evaluates the performance at frame k using the information about $|\widehat{TP}_k|$ and $|\widehat{FP}_k|$ as:

$$\text{FAR}_k = \frac{|\widehat{FP}_k|}{|\widehat{TP}_k| + |\widehat{FP}_k|}. \tag{2.16}$$

TDR provides the track-level tracking performance as follows:

$$\text{TDR}_i = \frac{|\widehat{TP}_j|}{\bar{K}_i}, \tag{2.17}$$

where $|\widehat{TP}_j|$ is the number of true positive targets in $\mathfrak{X}_j$ and $\bar{K}_i$ is the number of frames in the corresponding ground-truth track.

Track Fragmentation (TF) evaluates the ID consistency of targets in the form of their number of ID changes, $|IDC_i|$:

$$\text{TF}_i = |IDC_i|; \tag{2.18}$$

$|IDC_i|$ is measured with respect to a ground-truth track i ($\bar{\mathfrak{X}}_i$) as the number of estimated tracks that are associated to $\bar{\mathfrak{X}}_i$, where the association is established as for OTE.

2.3.5 Region-based assignment and position-based evaluation measures for multi-target tracking

RAPTE measures are true positive track matches, false positive track matches and false negative matches [13], as discussed below.

True positive (TP) track matches, false positive (FP) track matches and false negative (FN) track matches are determined using the spatial and temporal overlaps between estimated and ground-truth tracks. The computation of these measures involves solving the assignment problem implicitly. If there exists a spatial overlap as well as a temporal overlap between the estimated track j and any ground-truth track i, it is considered to be a TP track match. A spatial overlap is measured between the estimated track j and ground-truth track i as the percentage of frames in which the centroid of the estimated bounding box coincides with the corresponding ground-truth bounding box. In the case of TP track match, the temporal overlap, $\breve{O}_{tp}$, is computed as a ratio of the number of frames that are in common between estimated track j and ground-truth track i

($\mathcal{N}_{i,j}^{ov}$), and $\bar{K}_i$:

$$\breve{O}_{tp} = \frac{\mathcal{N}_{i,j}^{ov}}{\bar{K}_i}, \tag{2.19}$$

An estimation is a FP track match if the spatial or temporal overlap is less than a threshold, τ_3. In the case of FP track match, the temporal overlap, $\breve{O}_{fp}$, is defined as

$$\breve{O}_{fp} = \frac{\mathcal{N}_{i,j}^{ov}}{K_j}. \tag{2.20}$$

An estimation is a FN track match if the spatial or temporal overlap is less than a threshold, τ_4. In this case, the spatial overlap is measured between the ground-truth track i and any estimated track j as the percentage of frames in which the centroid of the ground truth bounding box coincides with the corresponding estimated bounding box. The temporal overlap for a FN track match is $\breve{O}_{fn} = \breve{O}_{tp}$ (Eq. 2.19).

2.3.6 Region-based assignment and size-based evaluation measures for multi-target tracking

RASTE measures are Correct Detected Track [117], False Alarm Track [117], Track Detection Failure [117], Multiple Object Tracking Precision [52], Multiple Object Detection Accuracy [52], Normalised Multiple Object Detection Accuracy [52], Multiple Object Tracking Accuracy [52] and ID changes [117], as explained below.

The concept of Correct Detected Track (CDT), False Alarm Track (FAT) and Track Detection Failure (TDF) is similar to TP, FP and FN track matches, respectively (Sec. 2.3.5). However, the difference is that the latter set of measures establish the spatial overlap in a frame based on the coincidence criterion and the former set of measures compute the spatial overlap based on the amount of common pixels between estimated and ground-truth target regions (as defined in Eq. 2.3). This means that, unlike TP, FP and FN track matches, CDT, FAT and TDF also incorporate target-size changes in the evaluation.

Multiple Object Tracking Precision (MOTP), Multiple Object Detection Accuracy (MODA), Normalised MODA (N-MODA) and Multiple Object Tracking Accuracy (MOTA) use a one-to-one frame-level assignment between estimated and ground-truth tracks determined by maximising the spatial overlap values (calculated as in Eq. 2.3) between track pairs using the Hungarian algorithm [56, 52].

MOTP evaluates the tracking performance by quantifying the amount of overlap between estimated and ground-truth tracks as follows:

$$\text{MOTP} = \frac{\sum_{i=1}^{n_m} \sum_{k=\hat{k}_{ini}^i}^{\hat{k}_{end}^i} \frac{|\bar{A}_k^i \cap A_k^i|}{|\bar{A}_k^i \cup A_k^i|}}{\sum_{k=1}^{K} n_m^k}, \tag{2.21}$$

where n_m denotes the number of associated pairs of estimated and ground-truth tracks, and n_m^k denotes the number of associated pairs of the estimated and ground-truth targets at frame k. MOTP uses in the evaluation the pairs with an overlap, $\frac{|\bar{A}_k^i \cap A_k^i|}{|\bar{A}_k^i \cup A_k^i|} > \tau_o$, where τ_o is a fixed threshold value.

MODA_k uses the information about the number of false negatives ($|\widehat{FN}_k|$) and the number of false positives ($|\widehat{FP}_k|$) at frame k to evaluate tracking performance:

$$\text{MODA}_k = 1 - \frac{c_1|\widehat{FN}_k| + c_2|\widehat{FP}_k|}{v_k}, \tag{2.22}$$

where c_1 and c_2 are application-dependent parameters fixed *a priori*. The estimations are classified into $\widehat{FP}_k$ and $\widehat{FN}_k$ by comparing the overlap between the associated pairs of estimated and ground-truth target regions with the threshold τ_o. MODA does not have a lower bound and continues to decrease with the increase in $|\widehat{FN}_k|$ and/or $|\widehat{FP}_k|$. N-MODA is the sequence-level formulation of MODA and is given as

$$\text{N-MODA} = 1 - \frac{\sum_{k=1}^{K}(c_1|\widehat{FN}_k| + c_2|\widehat{FP}_k|)}{\sum_{k=1}^{K} v_k}. \tag{2.23}$$

MOTA provides the tracking performance at sequence level and incorporates also the contribution of the number of ID switches ($|IDS_k|$) in addition to the contributions of $|\widehat{FP}_k|$ and $|\widehat{FN}_k|$. The parameters c_1, c_2 and c_3 (fixed *a priori*) determine the contributions of $|\widehat{FN}_k|$, $|\widehat{FP}_k|$ and $|IDS_k|$, respectively, which are combined over time and normalised as

$$\text{MOTA} = 1 - \frac{\sum_{k=1}^{K}(c_1|\widehat{FN}_k| + c_2|\widehat{FP}_k| + c_3|IDS_k|)}{\sum_{k=1}^{K} v_k}. \tag{2.24}$$

The computation of $\widehat{FP}_k$ and $\widehat{FN}_k$ is the same as in MODA. Moreover, like MODA, MOTA is numerically not lower bounded.

ID Changes (IDC) sums the number of ID changes belonging to all ground-truth tracks. The association between estimated and ground-truth bounding boxes is established at each frame by

Table 2.3: Summary of multi-target tracking evaluation measures. Key: PAPTE: Point-based Assignment and Position-based Tracking Evaluation; RAPTE: Region-based Assignment and Position-based Tracking Evaluation; RASTE: Region-based Assignment and Size-based Tracking Evaluation.

Measure	Ref.	Threshold independence	Size-change evaluation	Type	Evaluation	Accuracy computation	Cardinality error
OSPA	[86]			PAPTE	Frame level	✓	✓
$W_p(\cdot)$	[44]	✓		PAPTE	Frame level	✓	
OTE	[10]	✓		PAPTE	Sequence level	✓	
TRDR	[10]	✓		PAPTE	Frame level	✓	
FAR	[10]	✓		PAPTE	Frame level	✓	
TDR	[10]	✓		PAPTE	Sequence level	✓	
TF	[10]	✓		PAPTE	Sequence level		
TP matches	[13]			RAPTE	Sequence level	✓	
FP matches	[13]			RAPTE	Sequence level	✓	
FN matches	[13]			RAPTE	Sequence level	✓	
CDT	[117]		✓	RASTE	Sequence level	✓	
FAT	[117]		✓	RASTE	Sequence level	✓	
TDF	[117]		✓	RASTE	Sequence level	✓	
IDC	[117]		✓	RASTE	Sequence level		
MODA	[52]		✓	RASTE	Frame level	✓	✓
N-MODA	[52]		✓	RASTE	Sequence level	✓	✓
MOTA	[52]		✓	RASTE	Sequence level	✓	✓
MOTP	[52]		✓	RASTE	Sequence level	✓	
METE	[J1]	✓	✓	RASTE	Frame level	✓	✓
MELT	[J1]	✓	✓	RASTE	Sequence level	✓	
NIDC	[J1]	✓	✓	RASTE	Sequence level		

comparing their overlap with a fixed threshold. An ID change occurs at frame k if the overlap between an estimated and ground-truth track pair becomes less than the threshold. Table 2.3 summarises the existing multi-target tracking evaluation measures.

2.4 Evaluation campaigns and projects

Several campaigns and projects aided the evaluation of video surveillance including Context Aware Vision using Image-based Active Recognition project (CAVIAR)[2], Evaluation du Traitement et de l'Interpretation de Sequences vidEO (ETISEO)[3], Classification of Events, Activities and Relationships (CLEAR)[4], Performance Evaluation of Tracking and Surveillance (PETS)[5], imagery Library for Intelligent Detection Systems (i-LIDS)[6], and Visual Object Tracking (VOT) challenge[7]. Next we give an overview of these campaigns.

[2]http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. Accessed March 2014.
[3]http://www-sop.inria.fr/orion/ETISEO/index.htm. Accessed March 2014.
[4]http://www.clear-evaluation.org/. Accessed May 2011.
[5]http://www.cvg.rdg.ac.uk/slides/pets.html. Accessed March 2014.
[6]http://www.ilids.co.uk. Accessed March 2014.
[7]http://www.votchallenge.net/index.html. Accessed March 2014.

2.4.1 Context Aware Vision using Image-based Active Recognition

The Context Aware Vision using Image-based Active Recognition (CAVIAR) project was an important initiative which was mainly aimed at detection, activity recognition and behaviour analysis in indoor city surveillance scenarios. The project made contributions in the form of efficient feature extraction, activity and behavior recognition methods, performance comparison of detection methods (using TRDR (Eq. 2.15) and FAR (Eq. 2.16)), and numerous datasets for building entrance lobby and shopping mall scenes[8]. The building entrance sequences were recorded with two cameras (one with a wide-angle lens and the other with a steerable pan-tilt-zoom) and contain activities such as walking, browsing, resting, fainting, bag leaving, people meeting, people splitting and people fighting. The shopping mall sequences were recorded from two views (one with a corridor inside view and the other with a frontal view) and cover activities such as people walking along the corridor, people going in and out of stores, people crossing their paths, people stopping in the corridor and people browsing. These datasets are used by the community also for testing the tracking algorithms. Moreover, CAVIAR contributed and sponsored workshops including the Computer Vision System Control Architectures (VSCA)[9] in 2003, the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)[10] in 2004, and the Human Activity Recognition and Modelling (HAREM)[11] in 2005.

2.4.2 Evaluation du Traitement et de l'Interpretation de Sequences vidEO

Evaluation du Traitement et de l'Interpretation de Sequences vidEO (ETISEO) was a campaign aimed at evaluating the video surveillance algorithms for detection, tracking, classification and event recognition. The campaign provided a large number of datasets including sequences and ground truth, and introduced evaluation measures. The datasets used in ETISEO covered indoor and outdoor scenarios of the building corridor, building entrance, aircraft parking area, street/road, room, car park and metro with people and/or persons as targets. The activities included vehicle motion; person and object interactions; person, vehicle and object interactions; and abondoned baggage in crowded scenarios. The sequences were recorded using infrared and colour single and multiple cameras. The measures used in tracking evaluation quantified the percentage of frames where the target is tracked (a target was considered tracked in a frame if

[8] http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/. Accessed June 2012.
[9] http://homepages.inf.ed.ac.uk/rbf/VSCA03/. Accessed May 2014.
[10] http://www-prima.inrialpes.fr/PETS04/pets04.html. Accessed May 2014.
[11] http://users.isr.ist.utl.pt/~jasv/harem2005/. Accessed May 2014.

$D_k > 0.7$) and ID changes [73, 74]. The campaign contributed by disseminating data and conducting a series of meetings in which a total of 17 teams belonging to various universities and organisations participated.

2.4.3 Classification of Events, Activities and Relationships

Classification of Events, Activities and Relationships (CLEAR) was an important effort within the ambit of performance evaluation of surveillance systems. In particular, the campaign was aimed towards evaluating algorithms involving detection and tracking (person, face, vehicle), person identification, head pose estimation, and acoustic event detection and classification. The dataset for the CLEAR evaluation campaign was provided in collaboration with the Computers in the Human Interaction Loop (CHIL) project [106] and imagery Library for Intelligent Detection Systems (i-LIDS). The sequences covered scenarios including meeting room, lecture room and surveillance. The measures used in CLEAR included Multiple Object Detection Accuracy (MODA) (Eq. 2.22), Normalised-MODA (N-MODA) (Eq. 2.23), Multiple Object Tracking Accuracy (MOTA) (Eq. 2.24), Multiple Object Tracking Precision (MOTP) (Eq. 2.21). Moreover, CLEAR organised two workshops in 2006 and 2007.

2.4.4 Performance Evaluation of Tracking and Surveillance

Performance Evaluation of Tracking and Surveillance (PETS) was an evaluation program for video tracking and surveillance algorithms. Since 2000, PETS has organised a series of workshops in association with important international conferences. PETS provided the community with a wealth of key benchmark datasets to test their algorithms for a range tasks including outdoor tracking of people (PETS 2000[12], PETS 2001[13], PETS 2003[14]) and vehicle (PETS 2000, PETS 2001), indoor tracking of people (PETS 2002[15]), tracking in crowded scenes (PETS 2009[16], PETS 2013[17]), people counting (PETS 2002), classification of hand posture (PETS 2002), left luggage detection events (PETS 2006[5]), luggage theft detection events (PETS 2007[18]) and crowd analysis (PETS 2009, PETS 2013). These datasets were recorded from single and

[12] ftp://ftp.cs.rdg.ac.uk/pub/PETS2000/. Accessed June 2012.
[13] http://www.cvg.rdg.ac.uk/PETS2001/pets2001-dataset.html. Accessed March 2014.
[14] ftp://ftp.pets.rdg.ac.uk/pub/VS-PETS/. Accessed April 2014.
[15] ftp://ftp.pets.rdg.ac.uk/pub/PETS2002/. Accesed April 2014.
[16] http://www.cvg.rdg.ac.uk/PETS2009/. Accessed April 2014.
[17] http://pets2013.net/. Accessed April 2014.
[18] http://pets2007.net/. Accessed April 2014.

multiple views and involved key challenges. In PETS workshops the participants used various measures to evaluate results such as object centroid error, ID changes and overlap information.

2.4.5 Imagery Library for Intelligent Detection Systems

Another effort towards the evaluation of surveillance systems was made with the introduction of imagery Library for Intelligent Detection Systems (i-LIDS) that was created by the Centre for Applied Science and Technology (CAST)[19] in collaboration with the Centre for the Protection of National Infrastructure (CPNI)[20]. Various datasets were provided covering scenarios of the underground station, doorway surveillance, traffic (roads) and airport, which could be used to test the methods for event detection and tracking. i-LIDS provides real-world CCTV video footage for these scenarios making it useful for the evaluation of algorithms under challenging conditions. The evaluation criteria for measuring tracking performance involves classifying the frame-level estimations (bounding boxes) into true positives, false positives and false negatives based on their Euclidean distance from the ground truth. The total true positives, false positives and false negatives across the sequences are then used to quantify Precision (Eq. 2.7), Recall (Eq. 2.8) and F-score (Eq. 2.9).

2.4.6 Visual Object Tracking challenge

Visual Object Tracking (VOT) challenge was conducted in 2013 to aim at the evaluation of state-of-the-art single-target trackers using the existing commonly-used datasets[21]. The dataset covered scenarios such as diving, gymnastics, concert, sport, road/street, skating, jumping and indoor scenes. The tracking performance criteria used the extent of overlap (O_k (Eq. 2.3)) across the sequence. Specifically, two evaluation scores were used [54]. The first one was the mean overlap across the sequence to quantify the accuracy. The second one quantifies the number of times the tracker fails (i.e. $O_k = 0$) across the sequence. Indeed, this idea of separately quantifying accuracy and failure for evaluating trackers' performance is inspired from [J2], in which we compute tracking accuracy (using a procedure that uses temporal overlap information and lost-track ratio [62]) and failure (as a percentage of frames in the sequence with $O_k = 0$), and combine them into a single score.

[19] https://www.gov.uk/government/publications/introduction-to-the-centre-for-applied-science-and-technology. Accessed May 2014.

[20] http://www.cpni.gov.uk/. Accessed May 2014.

[21] http://box.vicos.si/vot/vot2013.zip. Accessed March 2014.

Despite the diffusion of several evaluation campaigns, there is still a lack of a commonly-used platform for the video tracking performance evaluation. VOT challenge aimed to address this shortcoming. It would also be desirable to enable an explicit robustness evaluation of trackers in the presence of distortions (that may influence the performance of trackers in real applications) including noisy inputs, frame dropping, video compression and varying scene conditions such as illumination changes, which would be needed when choosing a tracker for a specific challenge. Tab. 2.4 presents a summary of the evaluation campaigns and projects.

2.5 Datasets

Generally in video surveillance the trackers are evaluated on three types of target including head, person (full body) and vehicle. Below we list key sequences used by the community for the trackers' performance evaluation.

For *head* tracking evaluation the SPEVI Emilio[22], SPEVI Toni[22], Clemson[23], David [87], Occluded Face [2], Occluded Face 2 [3] and Sunshade[21] are well-known sequences (Fig. 2.7). All the sequences are captured in an indoor environment except Sunshade that is recorded outdoor. Among all the sequences David has significant illumination changes over time. Additionally, David and Clemson sequences involve constant camera motion. Moreover, SPEVI Emilio has significant variations in the target size and shows approximately a ten times increase in the target area compared to the initial target size. Furthermore, Occluded Face and Occluded Face 2 have a limited translatory motion of target but have recurring occlusions across the sequence. The specific challenges involved in these sequences include occlusions (SPEVI, Clemson, Occluded Face, Occluded Face 2), illumination changes (David, Sunshade), scale changes (SPEVI, Clemson, David), pose changes (SPEVI, Clemson, David), camera motion (Clemson, David, Occluded Face), clutter (Clemson, David, Occluded Face 2) and orientation changes (Sunshade, Occluded Face 2).

[22]http://www.eecs.qmul.ac.uk/~andrea/spevi.html. Accessed June 2012.
[23]http://www.ces.clemson.edu/~stb/research/headtracker/seq/. Accessed February 2014.

Table 2.4: Summary of the important evaluation campaigns and projects.

Name	Scope	Scenarios	Measure(s)
CAVIAR	Detection, activity recognition, behaviour analysis	Shopping mall, building entrance	TRDR, FAR
ETISEO	Detection, tracking, classification, event recognition	Building entrance/corridor, aircraft/car parking, room, metro	Dice, ID changes
CLEAR	Detection, tracking, person identification, head pose estimation, acoustic event detection and classification	Meeting room, lecture room, traffic	MODA, N-MODA, MOTA, MOTP
PETS	Tracking, people counting, event detection, crowd analysis	Road, traffic, metro, airport	Centroid error, ID changes, Overlap
iLIDS	Tracking, event detection	Metro, traffic, airport	Precision, Recall, F-score
VOT	Tracking	Gymnastics, diving, concert, sport, road, skating, jumping	Mean Overlap, no. of failures

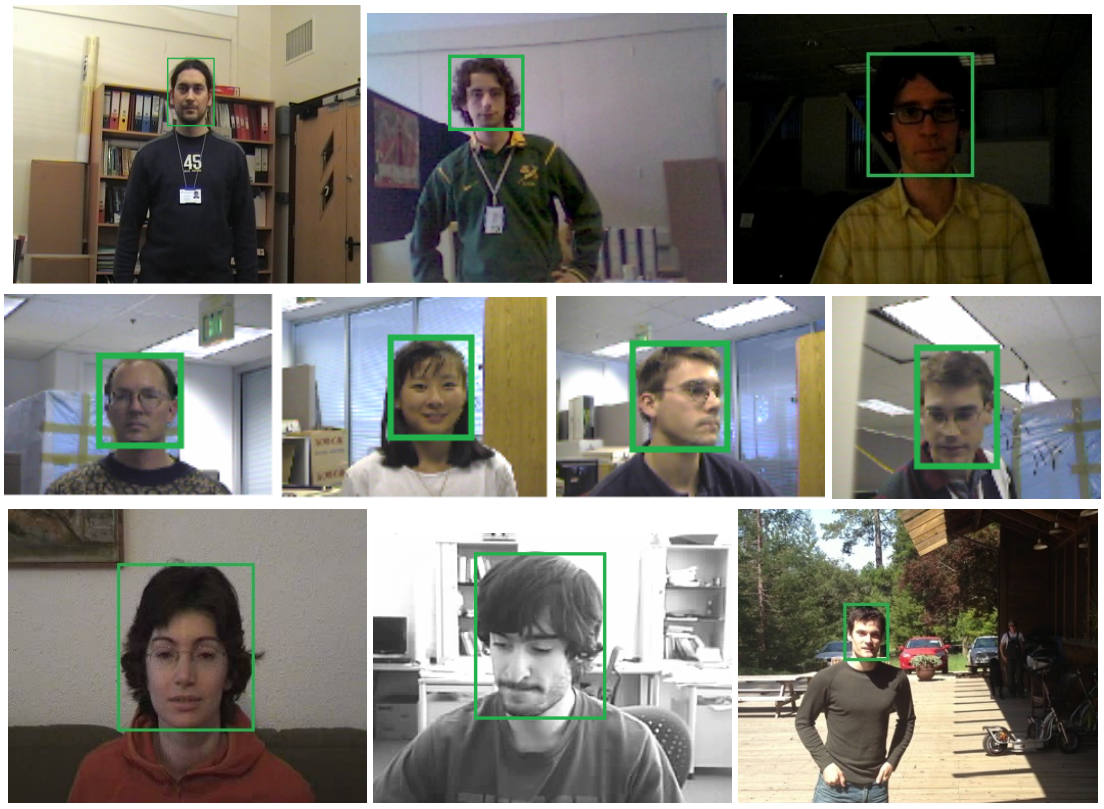


Figure 2.7: Samples from the datasets for the head tracking evaluation. First row (left to right): SPEVI Toni, SPEVI Emilio, David; second row: Clemson; third row (left to right) Occluded Face, Occluded Face 2, Sunshade.

For *person* tracking evaluation the ETH Bahnhof[24], ETH Sunnyday[24], TownCentre [8], iLids[25], PETS 2009[26], CAVIAR[8], Bolt[21], Diving[21], Gymnastics[21], Iceskater[21], Singer[21] and Woman [2] are important sequences (Fig. 2.8). TownCentre, ETH Bahnhof and ETH Sunnyday present crowded scenes with average number of people per frame of 16, 8 and 5, respectively. Singer involves significant target size changes and illumination changes across the sequence. Moreover, Bolt, Diving, Gymnatics and Ice skater have a highly articulate (fast) target motion (substantial pose and orientation changes) across the sequence. Specifically, the sequences include the challenges of occlusions (TownCentre, ETH Bahnhof, ETH Sunnyday, iLids, PETS 2009, CAVIAR, Bolt, Woman), crowdedness (TownCentre, ETH Bahnhof, ETH Sunnyday, Bolt), camera motion (ETH Bahnhof, ETH Sunnyday, Bolt, Diving, Gymnastics, Iceskater, Singer, Woman), illumination changes (ETH Bahnhof, ETH Sunnyday, iLids, Singer), scale changes (TownCentre, ETH Bahnhof, ETH Sunnyday, iLids, PETS 2009, CAVIAR, Bolt, Gymnastics, Iceskater, Singer,

[24]http://www.vision.ee.ethz.ch/~aess/iccv2007/. Accessed August 2012.
[25]http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html. Accessed June 2012.
[26]http://www.cvg.rdg.ac.uk/PETS2009/a.html#s2l1. Accessed May 2014.

Figure 2.8: Samples from the datasets for the person tracking evaluation. First row (left to right): ETH Bahnhof, ETH Sunnyday, TownCenter; second row (left to right): iLids, PETS 2009, CAVIAR, CAVIAR; third row (left to right): Bolt, Diving, Gymnastics; fourth row (left to right): Iceskater, Singer, Woman.

Woman), pose changes (PETS 2009, CAVIAR, Bolt, Gymnastics, Iceskater, Woman), variable speed (TownCentre, iLids, CAVIAR), clutter (PETS 2009, CAVIAR, Bolt, Diving, Gymnastics, Singer, Woman) and orientation changes (Diving, Gymnastics, Iceskater).

For *vehicle* tracking evaluation the PETS 2000[12], PETS 2001 Highway[13], PETS 2001 Reverse[13], AVSS 2007 challenge[25], Car [51], Bike [51], Bicycle[21] and Jump[21] are key sequences (Fig. 2.9). PETS 2001 Highway, Car, Bike, Bicycle and Jump involve a continuous target and camera motion across the sequence. PETS 2001 Reverse has a reverse and forward vehicle motion and 180^o pose change in the sequence. Moreover, AVSS 2007 challenge sequence has substantial scale changes and a strong background clutter. The specific challenges in the sequences include camera motion (Bicycle, PETS 2001, Jump, Car, Bike), occlusions (Bicycle, Car, Bike), clutter (Bicycle, PETS 2001, PETS 2000, AVSS 2007, Car, Bike), scale changes (Bicycle, Jump,

Figure 2.9: Samples from the datasets for the vehicle tracking evaluation. First row (left to right): PETS 2000, PETS 2001 Highway, PETS 2001 Reverse; second row: AVSS 2007, Car; third row (left to right) Bicycle, Jump, Bike.

PETS 2000, PETS 2001, AVSS 2007, Bike), pose changes (Bicycle, PETS 2000, PETS 2001, AVSS 2007, Jump), shadows (Jump, Car, Bike) and abrupt movement (Bike).

Tab. 2.5 lists the important sequences used for video tracking performance evaluation. For a comprehensive evaluation of trackers the choice of dataset should be made so as to consider the diversity of targets, presence of different challenges and the input variations accounting for the operational conditions in real-world applications, and an appropriate amount of test data to provide a statistically significant performance comparison of trackers. The choice of test sequences in this thesis is made taking into account these considerations and include sequences from PETS, CAVIAR, i-LIDS, ETH Bahnhof, ETH Sunnyday, SPEVI, TownCentre and Clemson datasets.

2.6 Discussion

We reviewed the state-of-the-art single-target tracking measures that provide distance-based or overlap-based evaluation (Tab. 2.2). Distance-based measures can be ineffective in identify-

Table 2.5: Summary of the key sequences for video tracking performance evaluation.

Sequence	Ref.	Target type	Challenges
SPEVI	[22]	Head	Occlusions, scale changes, pose changes
Clemson	[23]	Head	Occlusions, scale changes, pose changes, camera motion, clutter
David	[87]	Head	Illumination changes, scale changes, pose changes, camera motion, clutter
Occluded Face	[2]	Head	Occlusions, camera motion
Occluded Face 2	[3]	Head	Occlusions, clutter, orientation changes
Sunshade	[21]	Head	Illumination changes, orientation changes
ETH Bahnhof	[24]	Person	Occlusions, crowdedness, camera motion,illumination changes, scale changes
ETH Sunnyday	[24]	Person	Occlusions, crowdedness, camera motion, illumination changes, scale changes
TownCentre	[8]	Person	Occlusions, crowdedness, scale changes, variable speed
iLids	[25]	Person	Occlusions, illumination changes, scale changes, variable speed
PETS 2009	[26]	Person	Occlusions, scale changes, pose changes, clutter
CAVIAR	[8]	Person	Occlusions, scale changes, pose changes, variable speed, clutter
Bolt	[21]	Person	Occlusions, crowdedness, camera motion, scale changes, pose changes, clutter
Diving	[21]	Person	Camera motion, clutter, orientation changes
Gymnastics	[21]	Person	Camera motion, scale changes, pose changes, clutter, orientation changes
Iceskater	[21]	Person	Camera motion, scale changes, pose changes, orientation changes
Singer	[21]	Person	Camera motion, illumination changes, scale changes, clutter
Woman	[2]	Person	Occlusions, camera motion, scale changes, pose changes, clutter
PETS 2000	[12]	Vehicle	Clutter, scale changes, pose changes
PETS 2001	[13]	Vehicle	Camera motion, clutter, scale changes, pose changes
AVSS 2007 challenge	[25]	Vehicle	Clutter, scale changes, pose changes
Car	[51]	Vehicle	Camera motion, occlusions, clutter, shadows
Bike	[51]	Vehicle	Camera motion, occlusions, clutter, scale changes, shadows, abrupt movement
Bicycle	[21]	Vehicle	Camera motion, occlusions, clutter, scale changes, pose changes
Jump	[21]	Vehicle	Camera motion, scale changes, pose changes, shadows

ing tracking failures [95, 72, 67] and evaluating temporal size changes [72, 67], which makes the overlap-based measures more desirable. Existing overlap-based criteria [60, 62] use fixed threshold parameters that restrict their use to application-specific tracking performance assessment. Additionally, the measures [73, 60, 62] provide a frame-level tracking evaluation, whereas it would be desirable to effectively encapsulate the overall tracking performance into a single score to simplify the performance comparison task. Moreover, while several evaluation campaigns (CAVIAR, ETISEO, CLEAR, PETS, i-LIDS, VOT challenge), evaluation methods [60, 72, 73, 89] and datasets were introduced, researchers tend to use different measures and varying datasets to evaluate and compare their tracking algorithms due to the absence of a common platform unlike in other research areas of image processing and computer vision [47, 92, 5]. While the recent VOT challenge attempted to address this limitation, the need remains to enable the robustness evaluation of trackers in the presence of real-world conditions (under which trackers operate) using a uniform set of evaluation procedures.

We discussed the existing multi-target tracking evaluation measures, which are classified into different categories based on the type of their assignment and evaluation procedures (Tab. 2.3). Frame-level measures do not evaluate target-size variations [86, 10, 44], rely on preset application dependent thresholds [86, 52] and do not account for cardinality error [10, 44]. Likewise,

sequence-level measures do not take into account target-size-change evaluation (OTE, TDR [10] and those proposed in [13]), use preset thresholds (MOTA, MOTP [52] and those proposed in [117]) and provide accuracy evaluation only while ignoring cardinality error [10, 52, 117, 13]. Moreover, sequence-level measures are usually not used for analysing the performance of a tracker at varying accuracy levels, which would aid in determining its suitability for a specific scenario or application. Existing ID-change evaluation criteria simply count the number of ID changes or switches across the sequence [10, 52, 117], whereas it would be useful to quantify ID changes relative to the length of the track.

This thesis aims to address the above-mentioned limitations. Specifically, we propose three parameter-independent overlap-based measures to provide a comprehensive evaluation of multi-target tracking performance (Ch. 3). These measures are numerically bounded and take into account the temporal target-size changes. We also propose a protocol that uses a set of predefined evaluation procedures to assess and compare the robustness of trackers under a wide range of real-world operational conditions using a single-score evaluation criterion (Ch. 4). Finally, considering the lack of attention towards the assessment of tracking evaluation measures *per se*, we present a methodology to quantitatively assess the relative performance of measures (Ch. 5).

Chapter 3

Evaluation measures

3.1 Introduction

Discrepancy-based multi-target evaluation measures need to cope with the problem of establishing associations among estimated and ground-truth tracks followed by quantifying the tracking performance. The evaluation involves measuring accuracy of the estimate, target cardinality error and occurrence of ID changes (Fig. 2.2). Existing measures require parameter presetting [86, 52, 117], do not evaluate variations in target size over time [86, 44] and are not numerically bounded [52, 44], as highlighted in the previous chapter (Sec. 2.6). To address these limitations, we propose three measures [J1], namely Multiple Extended-target Tracking Error (METE), Multiple Extended-target Lost-Track ratio (MELT) and Normalised ID Changes (NIDC), for providing a thorough performance assessment of multi-target tracking taking into account accuracy, cardinality error and ID changes.

In this chapter, we first describe the three proposed measures, METE (Sec. 3.2), MELT (Sec. 3.3) and NIDC (Sec. 3.4). This is followed by the experimental validation and analysis in Sec. 3.5. The chapter is summarised in Sec. 3.6.

3.2 Multiple extended-target tracking error

Multiple Extended-target Tracking Error (METE) is an overlap-based measure that provides evaluation by combining accuracy and cardinality errors. Although inspired from OSPA (Eq. 2.13), METE is parameter-independent and does not require the inclusion of OSPA parameters (c, p)

due to the use of spatial overlap information.

The accuracy error, $\mathcal{A}_k$, computes the closeness between estimated and ground-truth states at frame k as

$$\mathcal{A}_k = \min_{\pi \in \Pi_{max(v_k,u_k)}} \sum_{i=1}^{min(v_k,u_k)} \left(1 - O(\bar{A}_{k,i}, A_{k,\pi(i)})\right), \tag{3.1}$$

where $O(\bar{A}_{k,i}, A_{k,\pi(i)})$ quantifies the extent of spatial overlap between the ground-truth target region, $\bar{A}_{k,i}$, and the estimated target region, $A_{k,\pi(i)}$: $O(\bar{A}_{k,i}, A_{k,\pi(i)}) = \frac{|\bar{A}_{k,i} \cap A_{k,\pi(i)}|}{|\bar{A}_{k,i} \cup A_{k,\pi(i)}|}$ as defined in Eq. 2.3. Without the loss of generality, $\bar{A}_{k,i}$ and $A_{k,\pi(i)}$ are considered to be bounding boxes. $\Pi_{max(v_k,u_k)}$ represents the set of permutations where $min(v_k,u_k)$ is the length of each permutation whose elements are taken from $\{1,2,\ldots,max(v_k,u_k)\}$. The permutation that minimises the summation term in Eq. 3.1 solves the assignment between estimated and ground-truth states, and is used to calculate the accuracy error, $\mathcal{A}_k$, at frame k. We use the Hungarian algorithm [56] for this minimisation. When $u_k = v_k$, $\mathcal{A}_k \in [0, u_k = v_k]$; when $u_k > v_k$, $\mathcal{A}_k \in [0, v_k]$ since the assignment is established only for the v_k terms; when $u_k < v_k$, $\mathcal{A}_k \in [0, u_k]$ since the assignment is established only for the u_k terms. As the computation of $\mathcal{A}_k$ considers only the associated pairs of estimated and ground-truth targets, the information about the unassociated targets (difference between u_k and v_k) needs also to be accounted for in the evaluation. This justifies the need of calculating the cardinality error, $\mathcal{C}_k$, which is defined as the error in estimating the number of targets:

$$\mathcal{C}_k = |u_k - v_k|. \tag{3.2}$$

Therefore, METE combines $\mathcal{C}_k$ with $\mathcal{A}_k$ (as done in the case of OSPA [86, 94]) to take into account the unassociated targets in the evaluation and to quantify the tracking performance into a single score at frame k as:

$$\text{METE}_k = \frac{\mathcal{A}_k + \mathcal{C}_k}{max(v_k, u_k)}, \tag{3.3}$$

$\text{METE}_k \in [0,1]$: the smaller METE_k, the better the tracking performance. Next we explain the bounded nature of the METE such that for the best tracking scenario $\text{METE}_k = 0$ and for the worst tracking scenario $\text{METE}_k = 1$.

Best tracking scenario: $O(\cdot) = 1$ for all associated pairs of the estimated and ground-truth targets, which implies $\mathcal{A}_k = 0$ (Eq. 3.1). $u_k = v_k$, which implies $\mathcal{C}_k = 0$ (Eq. 3.2). Therefore, $\text{METE}_k = 0$ using Eq. 3.3.

Worst tracking scenario: When $u_k = v_k$, $\mathcal{C}_k = 0$ and the highest value of $\mathcal{A}_k = u_k = v_k$; therefore the numerator in Eq. 3.3 is given as: $\mathcal{A}_k + \mathcal{C}_k = v_k = u_k$. When $u_k > v_k$, the highest value of $\mathcal{A}_k = v_k$; therefore the numerator becomes $\mathcal{A}_k + \mathcal{C}_k = v_k + |u_k - v_k| = u_k : u_k > v_k$. When $u_k < v_k$, the highest value of $\mathcal{A}_k = u_k$; therefore the numerator becomes $\mathcal{A}_k + \mathcal{C}_k = u_k + |u_k - v_k| = v_k : u_k < v_k$. Thus, the numerator boils down to $\mathcal{A}_k + \mathcal{C}_k = max(v_k, u_k)$ implying $\text{METE}_k = 1$ using Eq. 3.3. The other tracking cases lie between $\text{METE}_k = 0$ and $\text{METE}_k = 1$ (see Fig. 3.1).

The same METE values for two different trackers may result from different combinations of accuracy and cardinality errors. Therefore, separately analysing these errors may be useful to identify their individual contribution in the computation of METE. For this reason, we use the Accuracy Error Rate (AER):

$$\text{AER} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{A}_k, \tag{3.4}$$

and the Cardinality Error Rate (CER):

$$\text{CER} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{C}_k, \tag{3.5}$$

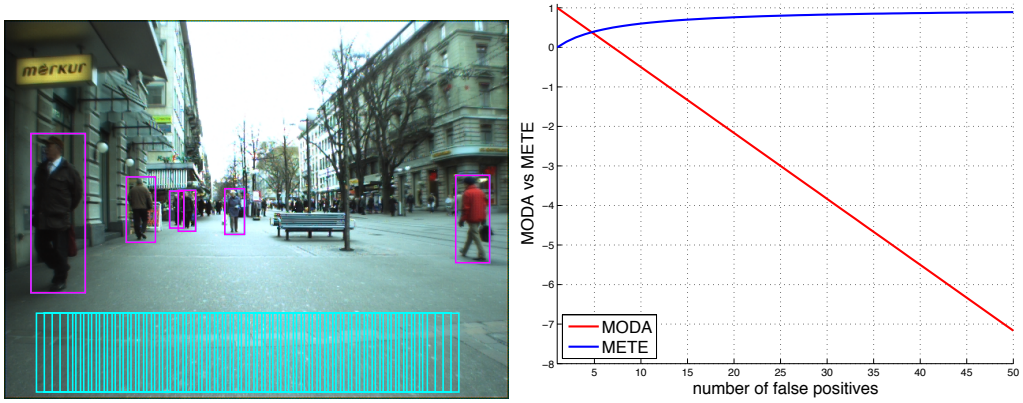
where K is the number of frames in the video sequence.

In summary, unlike OSPA, METE provides evaluation taking into account size changes of extended targets. Moreover, unlike MODA, METE has well-defined lower and upper bounds (Fig. 3.1) and is parameter independent. In fact, due to the parameter dependence, MODA may not always be able to distinguish different tracking results (Fig. 3.2).

3.3 Multiple extended-target lost-track ratio

The proposed Multiple Extended-target Lost-Track ratio (MELT) is a parameter-independent measure that provides sequence-level tracking accuracy evaluation and allows performance analysis at different accuracy levels. Given the sets of estimated ($\mathcal{X}$) and ground-truth tracks ($\bar{\mathcal{X}}$), the assignment is first solved at each frame by minimising the cost $1 - O(\cdot)$ (calculated for all estimated and ground-truth target pairs) using the Hungarian algorithm, as in Eq. 3.1. This results in a unique assignment at each frame. However, at track level a ground-truth track may be assigned to multiple estimated tracks because of ID changes and/or fragmentations.

For each associated pair of ground-truth track i and estimated track(s), the track-level accu-



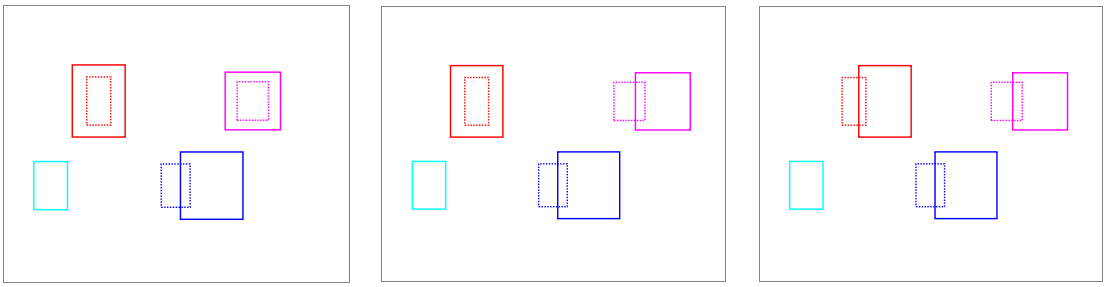
(a) (b)

Figure 3.1: An example to show that MODA is unbounded. (a) Sample frame is from ETH Bahnhof (six targets). A perfect overlap exists between the estimated bounding boxes (cyan) and corresponding ground-truth bounding boxes (magenta). The false positives (see the bottom of the frame) are gradually added from the perfectly overlapping scenario and the corresponding MODA and Multiple Extended-target Tracking Error (METE) values are computed and plotted in (b). While MODA is continuously decreasing (no lower bound), the proposed METE $\in [0,1]$.

racy is computed using the lost-track ratio (λ_i^τ) [62] as:

$$\lambda_i^\tau = \frac{N_i^\tau}{N_i}, \tag{3.6}$$

where N_i denotes the number of frames in the ground-truth track i and N_i^τ denotes the number of frames having $O(\cdot) \leq \tau : \tau \in [0,1]$. The spatial overlap $O(\cdot)$ is computed as in Eq. 2.3. $\lambda_i^\tau \in [0,1]$ such that the smaller λ_i^τ, the better the tracking performance. We compute λ_i^τ for a



MODA = -1.33, METE = 0.81 (a)

MODA = -1.33, METE = 0.88 (b)

MODA = -1.33, METE = 0.94 (c)

Figure 3.2: Illustration of a limitation of Multiple Object Detection Accuracy (MODA) [52]. MODA can not distinguish the three clearly different cases, whereas the proposed METE can do so. Ground-truth bounding boxes are shown as dotted lines and estimated bounding boxes are shown as solid lines. From (a) to (c) there is a gradual deterioration in the estimated results. Compared to (a), in (b) there is a deterioration (lesser overlap) in the magenta estimated bounding box only. Compared to (b), in (c) there is a deterioration (lesser overlap) in the red estimated bounding box only.

variation of τ values that yields $\lambda_i(\tau) = \{\lambda_i^\tau\}_{\tau\in[0,1]}$, where $\hat{S}_\tau$ denotes the number of τ values uniformly sampled from the range of $\tau = 0$ to $\tau = 1$. Likewise, $\lambda_i(\tau)$ is computed for all V ground-truth tracks to obtain the matrix

$$\Lambda = [\lambda_i^\tau]_{V\times\hat{S}_\tau}, \tag{3.7}$$

which has V rows and $\hat{S}_\tau$ columns. We therefore define tracking performance at τ using the Multiple Extended-target Lost-Track ratio (MELT_τ):

$$\text{MELT}_\tau = \frac{1}{V}\sum_{i=1}^{V}\lambda_i^\tau, \tag{3.8}$$

where $\text{MELT}_\tau \in [0,1]$: the smaller MELT_τ, the better the tracking performance. MELT_τ is computed for different τ values to enable performance analysis at different accuracy levels (Fig. 3.3(c), 3.3(d)), which may be useful from an application viewpoint. However, to simplify the task of performance comparison among trackers, we compute a single-score average tracking performance as

$$\text{MELT} = \frac{1}{\hat{S}_\tau}\sum_{\tau\in[0,1]}\text{MELT}_\tau. \tag{3.9}$$

The probability density function, H_τ, of lost-track-ratio values corresponding to a particular accuracy level, τ, (each column in the Λ-matrix (Eq. 3.7)) can be plotted to present the performance of a tracker (Fig. 3.3). The range of λ_i^τ ($\lambda_i^\tau \in [0,1]$) is divided into equal-width intervals (bins) to create the 'Bin' axis in Fig. 3.3. Each bin value of H_τ tells the percentage of tracks with the corresponding lost-track-ratio value at a specific τ. For an ideal tracking result the distribution of λ_i^τ values is concentrated towards bin zero at all τ values (Fig. 3.3(a)). Likewise, for a worst tracking result the distribution of λ_i^τ values is concentrated towards bin 1 at all τ values (Fig. 3.3(b)). Fig. 3.3(c) and Fig. 3.3(d) plot H_τ while varying τ for the results of the Conditional Random Field based tracker (CRFBT) [116] and the Dynamic Programming-Non-Maxima Suppression based tracker (DP-NMS) [81], respectively, on ETH Sunnyday dataset[1]. MELT=0.39 for CRFBT is better than MELT=0.56 for DP-NMS, which can also be noticed by the higher concentration of the distributions for CRFBT in the bins towards zero (Fig. 3.3(c)). The better performance of CRFBT can also be observed in its MELT_τ plot computed for a vari-

[1]http://www.vision.ee.ethz.ch/~aess/iccv2007/. Accessed August 2012.

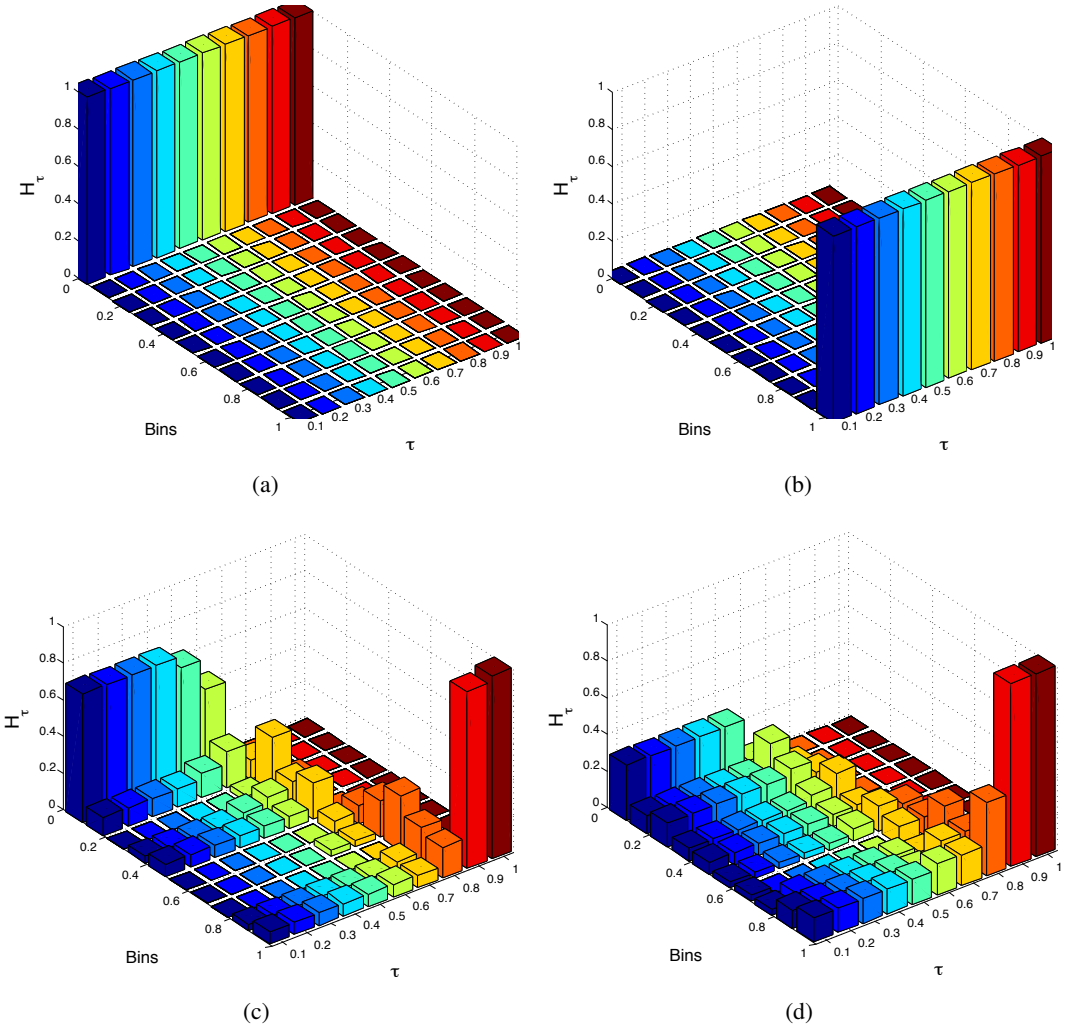


Figure 3.3: Examples of the probability density function (H_τ) plotted for a range of τ values. For an ideal tracking result (a) the distribution of lost-track ratio values is concentrated at bin zero because $\lambda_i^\tau = 0$ for all tracks at all τ values; therefore, MELT $= 0$. For a worst tracking result (b) the distribution of lost-track ratio values is concentrated at bin 1 because $\lambda_i^\tau = 1$ for all tracks at all τ values; therefore, MELT $= 1$. (c) and (d) show H_τ for Conditional Random Field based tracker (CRFBT) [116] and the Dynamic Programming-Non-Maxima Suppression based tracker (DP-NMS) [81] on ETH Sunnyday dataset, respectively, where MELT=0.39 and MOTP=0.75 for CRFBT, and MELT=0.56 and MOTP=0.77 for DP-NMS.

ation of τ (Fig. 3.4(c)). MELT$_\tau$ values are consistently lower (better) for CRFBT at all τ than for DP-NMS, hence a better tracking accuracy for the former. Conversely, MOTP ranks DP-NMS (MOTP=0.77) to be better than CRFBT (MOTP=0.75). This is because MOTP does not incorporate the complete tracking information in the evaluation procedure and considers only the overlap values of the estimated and ground-truth track pairs that are higher than τ_o (Eq. 2.21). MELT considers all of the tracking information thus providing a holistic tracking performance assessment.

MELT provides an insight into the performance by enabling the evaluation of tracking at

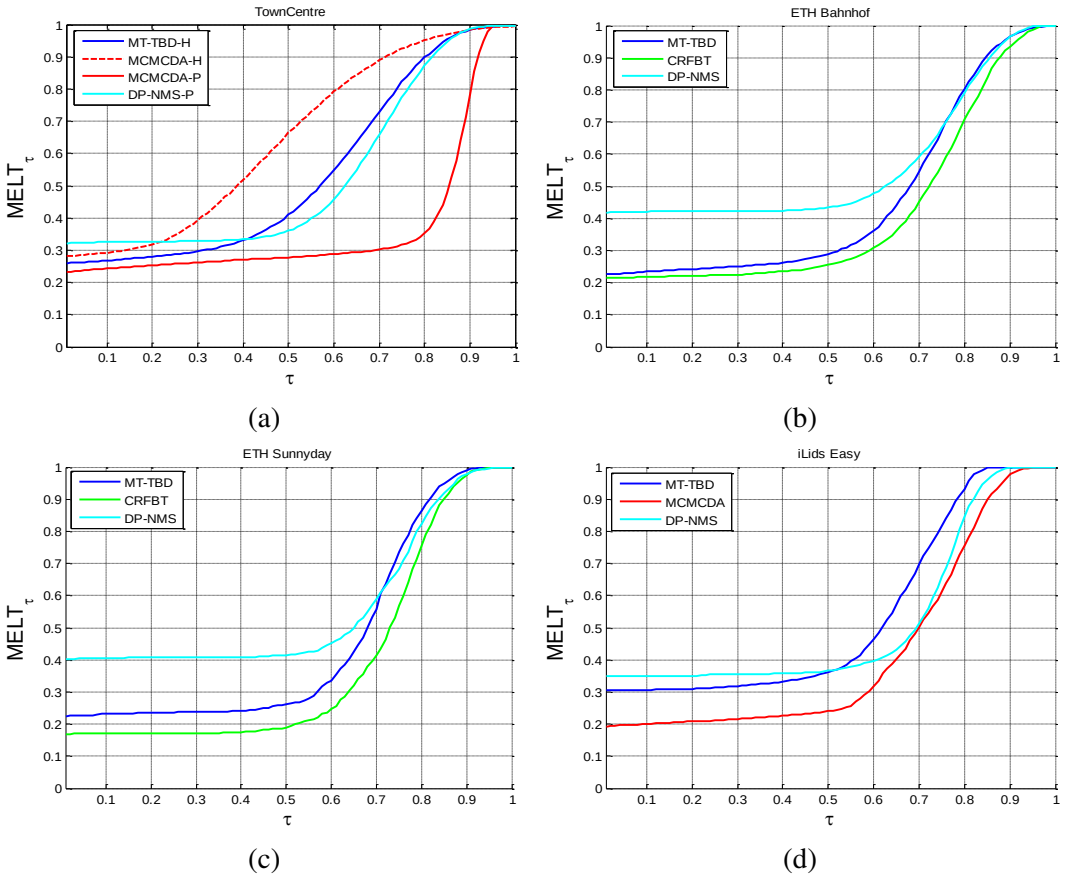


Figure 3.4: MELT_τ of trackers at varying accuracy levels (τ) on (a) TownCentre sequence with 'H' and 'P' in the legend referring to head and person tracking, respectively, (b) ETH Bahnhof sequence, (c) ETH Sunnyday sequence, and (d) iLids Easy sequence. Blue colour: MT-TBD; red colour: MCMCDA; cyan colour: DP-NMS; green colour: CRFBT.

different accuracy levels. Consider the MELT_τ plots of the multi-target track-before-detect (MT-TBD) tracker [83] and DP-NMS in Fig. 3.4(c). The performance of MT-TBD is better for $\tau < 0.72$ (approx.) and that of DP-NMS becomes better afterwards. Here MELT_τ can be helpful in choosing between MT-TBD and DP-NMS for an application with specific requirements. For example DP-NMS would be a more appropriate choice if the application requires tracking with an average overlap (accuracy) of 80%.

3.4 Normalised ID changes

The Normalised ID Changes (NIDC) measure provides the evaluation of the ID changes while accounting for the length of the track in which they occur. Compared to simply counting the number of ID changes, the normalisation by the length of the track is preferable when comparing trackers that generate tracks of different lengths. Such an evaluation enables the assessment of

the ability of trackers to track for long durations with unique IDs. Furthermore, the normalised score is preferable to the number of ID changes for the comparison of trackers across different datasets. Unlike MOTA [52] and IDC [117], NIDC does not require parameter presetting since the assignment procedure used for determining ID changes is based on the one employed in Sec. 3.2 (Eq. 3.1).

Let $|IDC_i|$ be the number of ID changes in the ground-truth track i and IDC_i^{max} be the maximum number of ID changes that can be produced for that track, we define the NIDC value for the ground-truth track i, NIDC_i, as follows:

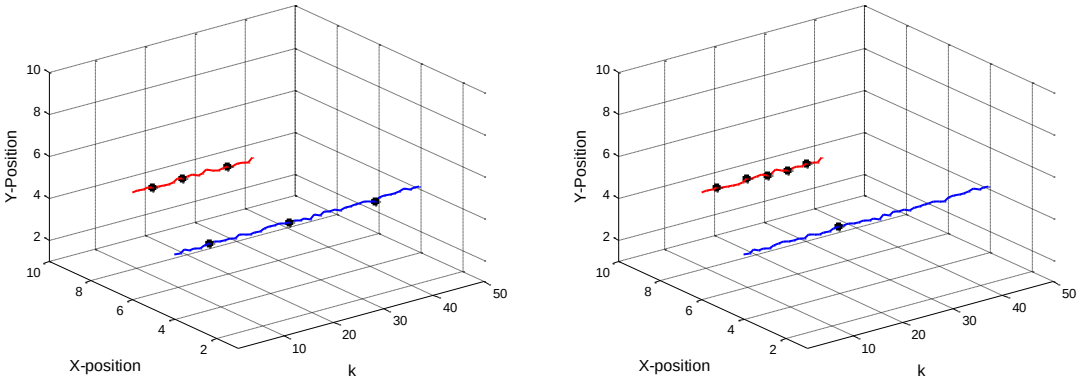
$$\text{NIDC}_i = \frac{|IDC_i|}{\text{IDC}_i^{max}}, \tag{3.10}$$

where the term IDC_i^{max} is (proportional to) the length of the track i. Therefore, the normalisation by IDC_i^{max} in NIDC_i (Eq. 3.10) penalises the ID changes by the duration of track in which they occur instead of simply counting the number of ID changes [10, 52, 117]. NIDC for the ID changes that have occurred for all ground-truth tracks of the sequence is defined as

$$\text{NIDC} = \frac{1}{V_{IDC}} \sum_{i=1}^{V} \text{NIDC}_i, \tag{3.11}$$

where V_{IDC} is the number of ground-truth tracks having ID change(s). $\text{NIDC} \in [0,1]$: the lower NIDC, the better the ID maintenance by the tracker.

We compare NIDC with Track Fragmentation (TF) [10] and ID Changes (IDC) [117] measures with examples (Fig. 3.5). Fig. 3.5(a) and Fig. 3.5(b) show ID changes for two different trackers on the same sequence, respectively. In Fig. 3.5(a), the number of ID changes for the red ground-truth track (ID=1) and the blue ground-truth track (ID=2) is the same ($|IDC_1| = |IDC_2| = 3$) but the length of the two tracks is different ($\text{IDC}_1^{max} = 25$, $\text{IDC}_2^{max} = 50$). $\text{NIDC}_1 = 0.12$ penalises the red track of shorter length more than the blue track with $\text{NIDC}_2 = 0.06$, despite the occurrence of the equal number of ID changes. On the other hand, $\text{TF}_1 = 3$ and $\text{TF}_2 = 3$ do not distinguish the two cases because the measure does not take into account the track length. NIDC and TF differentiate between ID changes of the red and blue tracks in Fig. 3.5(b) as reflected by their values. Furthermore, $\text{IDC} = 6$ for both cases (Fig. 3.5(a) and Fig. 3.5(b)) consider the two trackers to be the same, whereas NIDC distinguishes them ($\text{NIDC} = 0.09$ for Fig. 3.5(a) and $\text{NIDC} = 0.11$ for Fig. 3.5(b).



$\text{NIDC}_1 = 0.12$, $\text{TF}_1 = 3$
$\text{NIDC}_2 = 0.06$, $\text{TF}_2 = 3$
$\text{NIDC} = 0.09$, $\text{IDC} = 6$
(a)

$\text{NIDC}_1 = 0.20$, $\text{TF}_1 = 5$
$\text{NIDC}_2 = 0.02$, $\text{TF}_2 = 1$
$\text{NIDC} = 0.11$, $\text{IDC} = 6$
(b)

Figure 3.5: Comparison of the proposed NIDC with TF [10] and IDC [117]. Results are shown in (a) and (b) in terms of ID changes (black dots) for two different trackers on the same sequence. Both (a) and (b) show two ground-truth tracks. For the red ground-truth track ID=1 and for the blue ground-truth track ID=2. (a) The red track is shorter in length ($\text{IDC}_1^{max} = 25$) than the blue track ($\text{IDC}_2^{max} = 50$). Red track: $\text{NIDC}_1 = 0.12$; blue track: $\text{NIDC}_2 = 0.06$. The red track (shorter) is penalised for the same number of ID changes ($|IDC_1| = 3$) as the blue track ($|IDC_2| = 3$). $\text{TF}_1 = \text{TF}_2 = 3$ do not distinguish the two cases. (b) Both NIDC and TF can differentiate the two cases. $\text{IDC} = 6$ for both tracking cases ((a) and (b)) thus meaning that the measure considers both to be the same. NIDC distinguishes them (NIDC=0.09 for (a) and NIDC=0.11 for (b).

3.5 Experiments

We present the effectiveness of the proposed measures by showing their advantages over the existing measures and by evaluating and comparing the performance of recent trackers on real-world datasets. First we describe the datasets and trackers used in the experiments (Sec. 3.5.1). This is followed by discussing the advantages of the proposed measures (Sec. 3.5.2) and their use for evaluating trackers' performance (Sec. 3.5.3).

3.5.1 Datasets and trackers

We use four real-world publicly-available datasets, which include TownCentre [8], ETH Bahnhof[1], ETH Sunnyday[1] and iLids Easy[2]. The datasets are captured mostly in crowded scenes with occlusions. *TownCentre* is captured from an overhead static camera in Oxford. *ETH Bahnhof* and *ETH Sunnyday* are captured from a moving camera at a human height. *iLids Easy* is captured at the Westminster subway station in London. Tab. 3.1 presents a summary of the datasets.

[2]http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html. Accessed June 2012.

Table 3.1: Datasets used in the experiments. Key: K: number of frames; ANPF: average no. of people per frame.

Dataset	Frame size	K	Challenges	Camera	No. of tracks	ANPF	Trackers tested
TownCentre	1920×1080	4491	Occlusions, crowdedness, scale changes,variable speed	Static	231	16	[8, 81, 83]
ETH Bahnhof	640×480	999	Occlusions, crowdedness, illumination changes,scale changes	Moving	95	8	[81, 116, 83]
ETH Sunnyday	640×480	354	Occlusions, crowdedness, illumination changes,scale changes	Moving	30	5	[81, 116, 83]
iLids Easy	720×576	5220	Occlusions,illumination changes,scale changes,variable speed	Static	17	1.9	[8, 81, 83]

We use four state-of-the-art trackers in the experiments, which include (i) a method that combines Kanade-Lucas-Tomasi tracker [103] with a Markov-Chain Monte-Carlo Data Association (MCMCDA) algorithm [8], a data association technique with the online learned Conditional Random Field Based Tracker (CRFBT) [116], an algorithm based on Multi-Target Track-Before-Detect (MT-TBD) with a post-processing phase [83], and a method involving the Dynamic Programming Non-Maxima Suppression (DP-NMS) [81]. Tab. 3.1 lists the trackers used for each dataset. On TownCentre trackers are tested for head tracking (MT-TBD, MCMCDA) as well as person tracking (DP-NMS, MCMCDA), whereas on the remaining datasets trackers are tested for person tracking. The parameter values of trackers are the same as used in the original papers. To compute N-MODA for person tracking $\tau_o = 0.50$ and for head tracking $\tau_o = 0.25$, as used in [8].

3.5.2 Advantages of measures

We show the advantages of the proposed measures (METE, MELT, NIDC) by comparing them with the relevant state-of-the-art measures (N-MODA, MOTP, IDC). Tab. 3.2 lists the evaluation scores of trackers obtained using all the measures.

METE and N-MODA are in agreement with each other in their relative ranking of trackers on TownCentre with head tracking (TownCentre-H) and TownCentre with person tracking (TownCentre-P), and ETH Sunnyday. However, on ETH Bahnhof and iLids Easy the two measures show disagreements. For the case of Bahnhof, N-MODA of DP-NMS and MT-TBD is equal, which is because the measure uses the information about the number of false negatives and false positives (Eq. 2.23) while not accounting for the number of true positives. The total false negatives and false positives is comparable for DP-NMS (3525) and MT-TBD (3514), which make their N-MODA comparable. It is interesting to note that the number of true positives for DP-NMS (5030) is smaller than that of MT-TBD (6222). Therefore, METE considers MT-TBD to be better than DP-NMS because it implicitly considers true positives, false negatives and false positives in the evaluation procedure. For the case of iLids Easy, N-MODA considers

Table 3.2: Evaluation of trackers using different measures on different datasets. The measures include Multiple Extended-target Tracking Error (METE), Multiple Extended-target Lost Track ratio (MELT), Normalised ID Changes (NIDC), Accuracy Error Rate (AER), Cardinality Error Rate (CER), Mean Length of ground-truth Tracks with id change(s) (MLT), Normalised Multiple Object Detection Accuracy (N-MODA), Multiple Object Tracking Precision (MOTP), ID Changes (IDC). The colour of the cells show the performance of trackers such that the darker the colour of a cell, the better the tracking performance. Key: TownCentre-H: heads of persons (Head target) tracked on TownCentre dataset; TownCentre-P: full body of persons (Person target) tracked on TownCentre dataset; μ: mean of the measure scores across the sequence; σ: for METE it is the standard deviation of its scores across the sequence, and for AER and CER it is the standard deviation of accuracy error ($\mathcal{A}$) and cardinality error ($\mathcal{C}$) scores across the sequence, respectively.

Tracker	Dataset	METE $\mu(\sigma)$	MELT	NIDC	AER (σ)	CER (σ)	N-MODA	MOTP	IDC	MLT
MT-TBD [83]	TownCentre-H	0.53 (0.08)	0.54	0.031	6.82 (2.54)	2.14 (1.92)	0.55	0.64	1798	320.00
MCMCDA [8]		0.62 (0.07)	0.65	0.038	8.48 (2.74)	1.82 (1.62)	0.48	0.52	1913	330.13
DP-NMS [81]	TownCentre-P	0.48 (0.08)	0.53	0.043	5.06 (1.52)	2.67 (2.02)	0.58	0.71	2637	321.61
MCMCDA [8]		0.33 (0.09)	0.37	0.030	3.64 (1.54)	1.81 (1.62)	0.62	0.86	1519	336.44
DP-NMS [81]	ETH Bahnhof	0.53 (0.13)	0.57	0.039	1.45 (0.69)	3.07 (1.85)	0.58	0.75	229	109.92
MT-TBD [83]		0.44 (0.12)	0.46	0.050	2.42 (1.19)	1.56 (1.34)	0.58	0.75	307	103.51
CRFBT [116]		0.39 (0.12)	0.42	0.035	1.99 (0.86)	1.49 (1.26)	0.68	0.77	158	124.91
DP-NMS [81]	ETH Sunnyday	0.44 (0.11)	0.56	0.042	1.16 (0.55)	1.34 (0.93)	0.66	0.77	43	68.68
MT-TBD [83]		0.47 (0.11)	0.46	0.041	1.60 (0.57)	1.09 (0.84)	0.61	0.73	56	91.50
CRFBT[116]		0.46 (0.12)	0.39	0.028	1.46 (0.52)	1.06 (0.78)	0.63	0.75	31	82.20
DP-NMS [81]	iLids Easy	0.40 (0.26)	0.52	0.011	0.40 (0.34)	0.65 (0.86)	0.60	0.74	105	632.87
MT-TBD [83]		0.53 (0.22)	0.54	0.007	0.50 (0.36)	0.96 (1.10)	0.63	0.70	54	632.87
MCMCDA [8]		0.36 (0.24)	0.43	0.029	0.51 (0.45)	0.50 (0.76)	0.62	0.75	227	605.06

MT-TBD to be the best because the number of its false negatives and false positives (3640) is lower than that of MCMCDA (3698) and DP-NMS (3843). METE provides a more effective performance evaluation taking into account also the true positives and considers MCMCDA to be the best since its true positives (7969) are greater than that of MT-TBD (6706) and DP-NMS (6632).

MELT and MOTP show agreement in terms of ranking the trackers on TownCentre-H and TownCentre-P (Tab. 3.2). However, they have disagreements on ETH Bahnhof, ETH Sunnyday and iLids Easy. On Bahnhof, MOTP is the same for MT-TBD and DP-NMS, whereas MELT considers MT-TBD to be better than DP-NMS. This difference in performance is also shown by the mostly better MELT_τ of MT-TBD than DP-NMS for the variation of τ in Fig. 3.4(b). The reason for the disagreement of MOTP is because it takes into account only the overlap values of (estimated and ground-truth) pairs that are greater than the threshold value τ_o, which may result in the exclusion of some of the tracking information in the evaluation. MELT considers all of the tracking information in the evaluation to provide a comprehensive performance assessment thus more effectively comparing the trackers. On ETH Sunnyday, the inconsistency between MELT and MOTP in terms of their ranking of trackers is already discussed in Sec. 3.3. On iLids

Easy, MELT scores and MELT_τ plots (Fig. 3.4(d)) of DP-NMS and MCMCDA show a clear improvement in the performance of the latter, whereas MOTP is comparable for the two trackers due to its dependence on fixed τ_o.

NIDC and IDC show agreement in their relative ranking of trackers on all datasets except ETH Sunnyday (Tab. 3.2). Indeed, the usefulness of NIDC can be highlighted on ETH Sunnyday. While IDC ranks DP-NMS as better than MT-TBD, NIDC shows a slight improvement in the performance for MT-TBD as compared to DP-NMS despite the more ID changes of the former than the latter. The reason is that NIDC evaluates ID changes while taking into account also the length of tracks. Since MT-TBD has a higher MLT (mean length of the ground-truth tracks with ID change(s)) than DP-NMS, NIDC penalises its ID changes lesser than that of DP-NMS.

To summarise, MODA has a limited ability to distinguish different tracking results due to its dependence on the overlap threshold (τ_o) (Fig. 3.6(a-c)), whereas METE is parameter independent and better distinguishes different results. Additionally, the dependence of MOTP on τ_o may lead to an inaccurate performance evaluation of tracking (Fig. 3.7), whereas MELT is parameter independent and provides a more effective performance evaluation.

3.5.3 Performance comparison of trackers

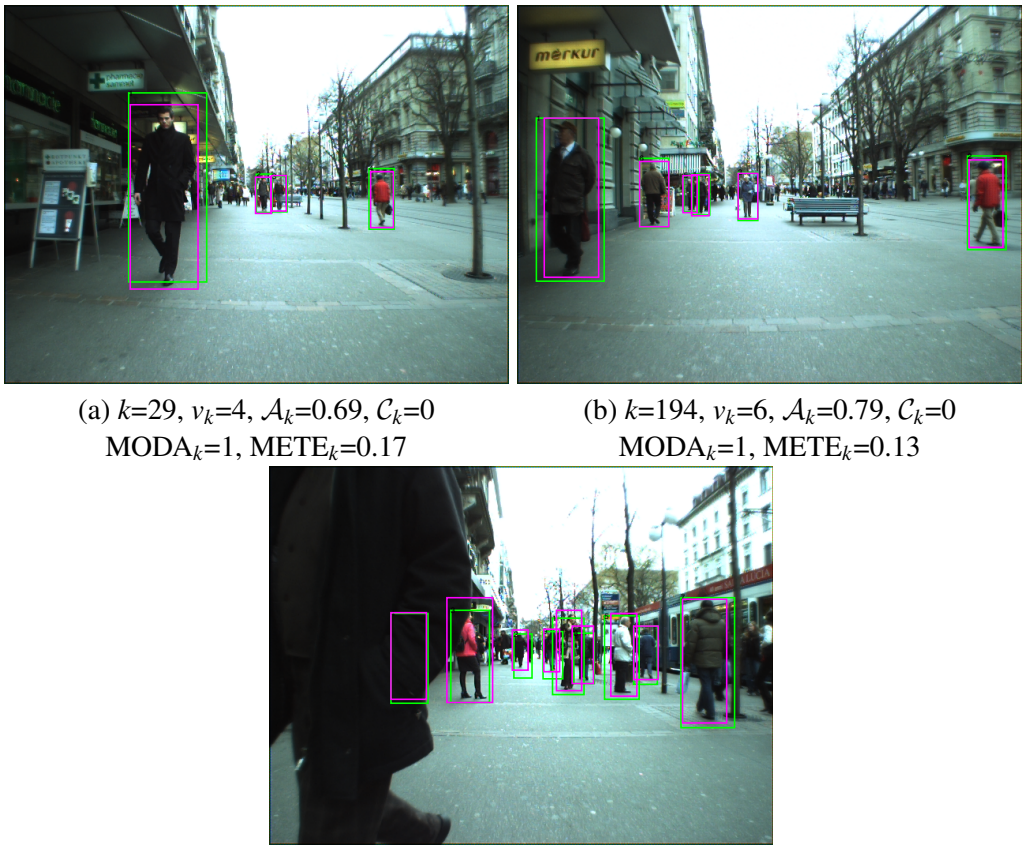
We show the usefulness of the proposed measures by comparing the performance of the selected trackers and discussing their strengths and limitations.

TownCentre-H: The mean METE, MELT, NIDC and AER show better performance of MT-TBD than MCMCDA (see r.[3] 1, 2 in Tab. 3.2). The better NIDC of MT-TBD is due to its more efficient ID management mechanism, which enables it to handle the mixing of particles of nearby targets in the Particle-filter-based state estimation framework [83]. Unlike the remaining measures, CER declares the performance of MCMCDA to be better than MT-TBD due to more occurrences of missed targets.

TownCentre-P: The mean METE, MELT, NIDC, AER and CER of MCMCDA are better than DP-NMS (see r. 3, 4 in Tab. 3.2). Interestingly, compared to TownCentre-H a clear improvement can be noticed in the performance of MCMCDA on TownCentre-P based on the proposed measures, which corresponds to the findings of the original paper [8].

ETH Bahnhof: CRFBT outperforms DP-NMS and MT-TBD based on mean METE, MELT,

[3] 'r' points to the row number (in Tab. 3.2 while skipping the first row with titles) that is under consideration.



(a) k=29, v_k=4, $\mathcal{A}_k$=0.69, $\mathcal{C}_k$=0
MODA$_k$=1, METE$_k$=0.17

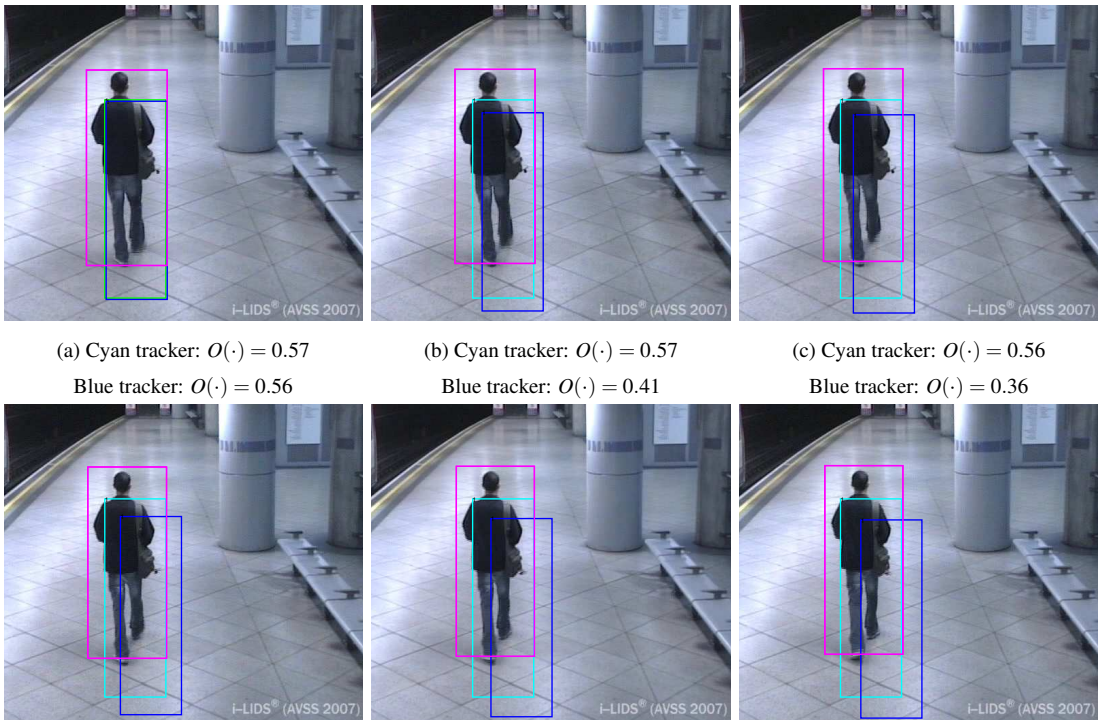
(b) k=194, v_k=6, $\mathcal{A}_k$=0.79, $\mathcal{C}_k$=0
MODA$_k$=1, METE$_k$=0.13

(c) k=616, v_k=9, $\mathcal{A}_k$=2.27, $\mathcal{C}_k$=0
MODA$_k$=1, METE$_k$=0.25

Figure 3.6: Sample results of CRFBT on Bahnhof with METE and MODA values listed for each case. Magenta colour: ground truth; green colour: tracker's result.

NIDC and CER (see r. 5, 6, 7 in Tab. 3.2). CRFBT has the best NIDC since it effectively handles ID changes due to the use of the motion and appearance 'affinities' [116]. Based on AER, DP-NMS outperforms CRFBT, which is not consistent with the remaining measures. Moreover, DP-NMS has a much higher CER than MT-TBD and CRFBT, which is due to its lesser capability to link fragmented tracks (thus increasing the cardinality error) that can be caused due to long-term occlusions (Fig. 3.8).

ETH Sunnyday: As on ETH Bahnhof, we evaluated DP-NMS, MT-TBD and CRFBT on this dataset. The following inconsistencies can be noticed in the results on ETH Sunnyday than on ETH Bahnhof. First, mean METE ranks the performance of DP-NMS to be better than MT-TBD and CRFBT. This ranking is not consistent with the ranking of trackers on ETH Bahnhof (see r. 8, 9, 10 in Tab. 3.2). The better performance of DP-NMS on ETH Sunnyday is probably due to its person detector [33], which can better handle the increased scene brightness in the scene as compared to the detector [111] in MT-TBD and CRFBT. Second, inconsistent with the



(a) Cyan tracker: $O(\cdot) = 0.57$
Blue tracker: $O(\cdot) = 0.56$

(b) Cyan tracker: $O(\cdot) = 0.57$
Blue tracker: $O(\cdot) = 0.41$

(c) Cyan tracker: $O(\cdot) = 0.56$
Blue tracker: $O(\cdot) = 0.36$

(d) Cyan: $O(\cdot) = 0.56$
Blue: $O(\cdot) = 0.32$

(e) Cyan: $O(\cdot) = 0.55$
Blue: $O(\cdot) = 0.29$

(f) Cyan: $O(\cdot) = 0.55$
Blue: $O(\cdot) = 0.27$

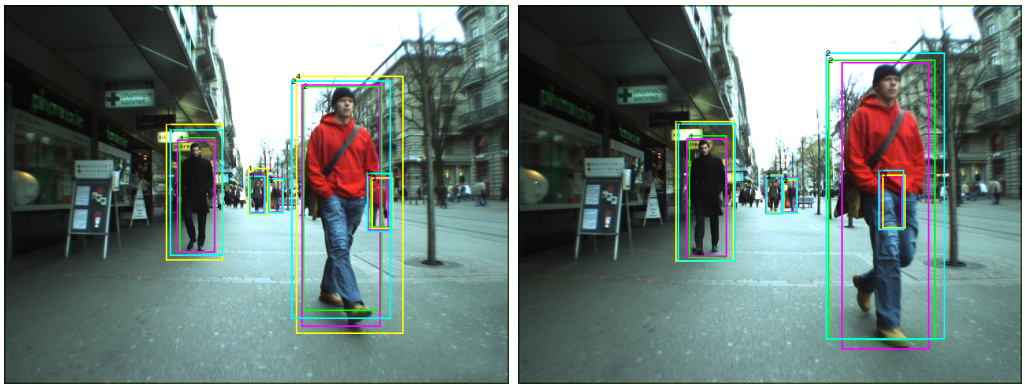
Figure 3.7: Due to parameter dependence MOTP [52] does not differentiate between two tracking results, whereas MELT does. MOTP=0.56 and MELT=0.45 for the case of Cyan tracker; MOTP=0.56 and MELT=0.64 for the case of Blue tracker. Ground truth is shown in magenta colour.

results of ETH Bahnhof, MT-TBD has a better NIDC than DP-NMS on ETH Sunnyday despite the higher IDC of the former due to the reason highlighted in Sec. 3.5.2.

iLids Easy: MCMCDA is the best based on mean METE and MELT (see r. 11, 12, 13 in Tab. 3.2). Like on TownCentre-P, MCMCDA has a better mean METE and MELT than DP-NMS on iLids Easy. Additionally, it is interesting to note from Fig. 3.4(d) that MT-TBD has a better MELT_τ compared to DP-NMS for $\tau < 0.5$ and for the remaining variation of τ DP-NMS performs better than MT-TBD, suggesting the use of DP-NMS when tracking with higher accuracy is desirable. Furthermore, CER of MT-TBD is the highest on iLids Easy, which is not consistent with the remaining datasets where DP-NMS has the highest CER. Moreover, the better ID management helped MT-TBD to obtain the best NIDC on iLids Easy. Despite the equal MLT of DP-NMS and MT-TBD[4] (see r. 11, 12 in Tab. 3.2), NIDC is lower (better) for the former due to its smaller IDC.

Table 3.2 also shows the performance variation of trackers in the form of standard deviation

[4]MLT is the same for the two trackers due to the occurrence of ID change(s) in the same tracks.



(a) $k = 16$ (b) $k = 19$

(c) $k = 24$

Figure 3.8: Results of trackers on Bahnhof for the case of a target occlusion. Green colour: DP-NMS; Yellow colour: MT-TBD; cyan colour: CRFBT; magenta colour: ground truth. Because of the occlusion DP-NMS misses the target in (b), whereas MT-TBD and CRFBT successfully deal with it.

(σ) values for different measures. As for METE, the σ values of trackers are comparable on all datasets. In the case of AER, the performance variation (σ) of MCMCDA is the highest on TownCentre and iLids Easy, whereas σ of MT-TBD is the highest on Bahnhof and Sunnyday. In the case of CER, on each dataset the σ values follow the same trend as the corresponding CER values of trackers.

To conclude, the results show that MCMCDA is better at person tracking than head tracking. DP-NMS has mostly shown the best accuracy but the highest cardinality error. A reason of the highest cardinality error of DP-NMS could be its lesser ability to cope with occlusions. Moreover, CRFBT is the best in terms of the ID change evaluation, followed by MT-TBD.

3.6 Summary

In this chapter, we introduced three evaluation measures, namely METE, MELT and NIDC, to account for the main aspects of extended multi-target tracking that are accuracy, cardinality error

and ID changes. The proposed measures are numerically bounded, parameter independent and take into account the temporal variations in target size. METE effectively combines accuracy and cardinality errors to provide a holistic performance assessment. MELT provides a deeper insight into the tracking performance by enabling analysis at varying accuracy levels, which can help in choosing trackers for particular applications. NIDC evaluates by normalising the ID changes with the duration of the track where they occur. We performed extensive experimentation to validate and compare the proposed measures with the existing measures using challenging real-world datasets and state-of-the-art multi-target trackers. We provided the source code for the measures online (http://www.eecs.qmul.ac.uk/~andrea/mtte.html) to facilitate their use for the community.

The focus of this chapter was to investigate effective ways to quantify the estimated results of trackers without taking into account their robustness in the presence of various operational conditions that they would need to cope with in real-world applications. These conditions represent distortions including noisy inputs, illumination changes, frame dropping, and video compression. To this end we present in the next chapter an evaluation protocol with a set of pre-defined procedures to enable testing the robustness of trackers under these conditions and show its effectiveness in the context of single-target tracking.

Chapter 4

Evaluation protocol

4.1 Introduction

Trackers operate under various conditions in real-world applications. The conditions allude to the introduction of distortions to the input of a tracker that can influence tracking performance, and may include initialisation errors caused by a detector, sensor noise, frame dropping caused during the transmission of video data over a channel or due to the delayed generation of results by the tracker, changing illumination in the scene, and compression of the video data (Fig. 4.1). Therefore, for testing the suitability of trackers at coping with these conditions, the tracking performance should be evaluated and compared in the presence of different distortions. To this end, we propose an evaluation protocol with a set of trials that test trackers on several test scenarios representing the above-mentioned real-world operational conditions [J2]. Specifically, the trials evaluate the *robustness* of a tracker under these distortions. Additionally, to quantify the trackers' results on trials we propose an overlap-based criterion (Combined Tracking Performance Score (CoTPS)) [J2] that summarises the overall tracking performance into a single score, which would simplify the performance comparison task on trials.

In this chapter we first define the problem in Sec. 4.2. The trials are explained in Sec. 4.3 on which the performance of trackers is quantified using the evaluation criterion described in Sec. 4.4. This is followed by experimental validation and analysis in Sec. 4.5, and the summary of the chapter in Sec. 4.6.

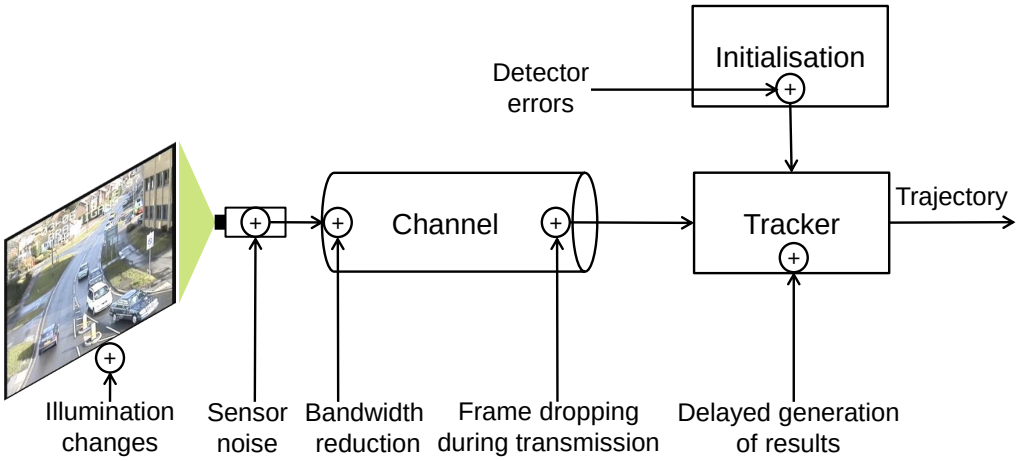


Figure 4.1: Conceptual illustration of various distortions that may affect the performance of a tracker in a real-world application. These distortions include initialisation errors caused by detector, sensor noise, frame dropping while transmitting video data over the channel or due to the delayed generation of results by the tracker, illumination changes in the scene, and video compression.

4.2 Problem definition

Given a set of $\mathcal{M}$ trackers[1] $\mathbf{T} = \{T_{\hat{j}}\}_{\hat{j}=1}^{\mathcal{M}}$, we aim to evaluate tracker $T_{\hat{j}}$ on a set of trials $\mathbf{P} = \{P_{\hat{i}}\}_{\hat{i}=1}^{Z} : Z = 8$, where trials simulate different real-world operational conditions (Fig. 4.1). On each trial $P_{\hat{i}}$, $T_{\hat{j}}$ is tested with the original (ground truth) initialisation I_t and the original video sequence $\mathcal{V}_t$ which contains a target h_t, where $\mathbf{h} = \{h_t\}_{t=1}^{J}$ is a set of targets. To study its variation in performance, each tracker $T_{\hat{j}}$ is tested with the initialisation $I_{t,\hat{i}}$ and test sequence $\mathcal{V}_{t,\hat{i}}$ which are generated on trial $P_{\hat{i}}$ by modifying I_t or $\mathcal{V}_t$ in a pre-defined manner such that the applied modification simulates a specific real-world scenario: $I_{t,\hat{i}} = P_{\hat{i}}(I_t)$ and $\mathcal{V}_{t,\hat{i}} = P_{\hat{i}}(\mathcal{V}_t)$.

Let $\mathfrak{X}_{t,\hat{i}}^{\hat{j}}$ be the trajectory of target h_t estimated by testing tracker $T_{\hat{j}}$ on trial $P_{\hat{i}}$ with $\mathcal{V}_{t,\hat{i}}$ and $I_{t,\hat{i}}$: $\mathfrak{X}_{t,\hat{i}}^{\hat{j}} = T_{\hat{j}}(\mathcal{V}_{t,\hat{i}}, I_{t,\hat{i}})$. The performance of tracker $T_{\hat{j}}$ is computed by evaluating the estimated trajectory $\mathfrak{X}_{t,\hat{i}}^{\hat{j}}$ of the target with respect to its ground-truth (ideal) trajectory $\bar{\mathfrak{X}}_t$ using the proposed evaluation criterion (Combined Tracking Performance Score (CoTPS)) thus obtaining the performance score: $\text{CoTPS}_{t,\hat{i}}^{\hat{j}} = \bar{\Psi}(\mathfrak{X}_{t,\hat{i}}^{\hat{j}}, \bar{\mathfrak{X}}_t)$, where $\bar{\Psi}(\cdot)$ represents the procedure involved in the proposed evaluation criterion. Based on $\text{CoTPS}_{t,\hat{i}}^{\hat{j}}$, we compare the performance of the trackers under consideration.

Next we describe the proposed trials (Sec. 4.3) on which trackers' performance can be quantified using the proposed evaluation criterion (Sec. 4.4).

[1]Please note that the index $\hat{j}$ refers to different trackers or to different parameter settings for the same tracker.

4.3 Trials

Trials 1, 2, 3 (P_1, P_2, P_3) evaluate the robustness of trackers to *initialisation errors* possibly introduced by a detector. These errors are simulated by perturbing the position of the initialising bounding box in *Trial 1* (P_1), the size (width and height) of the bounding box in *Trial 2* (P_2), and both the position and size in *Trial 3* (P_3). The amount of perturbation is added while ensuring at least an overlap of $\hat{O}\%$ between the bounding boxes of the original (ground-truth) initialisation and the perturbed initialisation. The number of perturbed initialisations generated on P_1, P_2 and P_3 are n_1, n_2 and n_3, respectively.

Trial 4 (P_4) evaluates robustness to *noisy video data* generated by low-cost sensors. On P_4, a set of n_4 test sequences are generated by adding to the original sequence $\hat{l}$ times (the estimated variance of) the zero-mean Gaussian noise of a low-quality webcamera (Creative webcam VF0330). The standard deviations of its red, green and blue channels are $\sigma_1 = 8.59$, $\sigma_2 = 8.40$ and $\sigma_3 = 11.96$, respectively. They are estimated by taking a pixel-based frame difference across a recorded sequence using the webcamera with fixed position under constant illumination in a static scene (Fig. 4.2), and modelling the distributions of the noise (difference between pixel values) in the three channels as Gaussians.

Trial 5 (P_5) evaluates robustness to cope with *frame dropping* that can be caused during the transmission of video data over a channel or due to the delayed generation of the results by the tracker. On P_5, the protocol generates a set of n_5 test sequences by periodically dropping $\hat{m}-1$ frames from the original sequence.

Trial 6 (P_6) evaluates robustness to *changing illumination* in the scene. On P_6, a set of n_6 test sequences is generated by synthetically increasing ($+\Delta L$) or decreasing ($-\Delta L$) illumination over time (in the original sequence) with saturation by adding (subtracting) $\Delta L = 0, 1, \ldots, L$ to (from) the pixel values of frames $k = 1, 2, \ldots, K$, respectively. If the number of frames in the sequence is $K > (L+1)$, a value of $\Delta L = L$ is maintained for the remaining frames.

Trials 7, 8 (P_7, P_8) evaluate robustness to *bandwidth reduction* of the video data. On P_7, test sequences are generated by gradually increasing the compression ratio of the original sequence. We chose Motion JPEG compression because of its suitability for video tracking applications [26]. In Motion JPEG, the extent of compression ratio depends on a quality parameter ζ. The higher ζ, the better the visual quality and the lower the compression ratio, where $\zeta \in [0, 100]$. To ensure evaluation under strong compression ratios, a set of n_7 test sequences are generated on

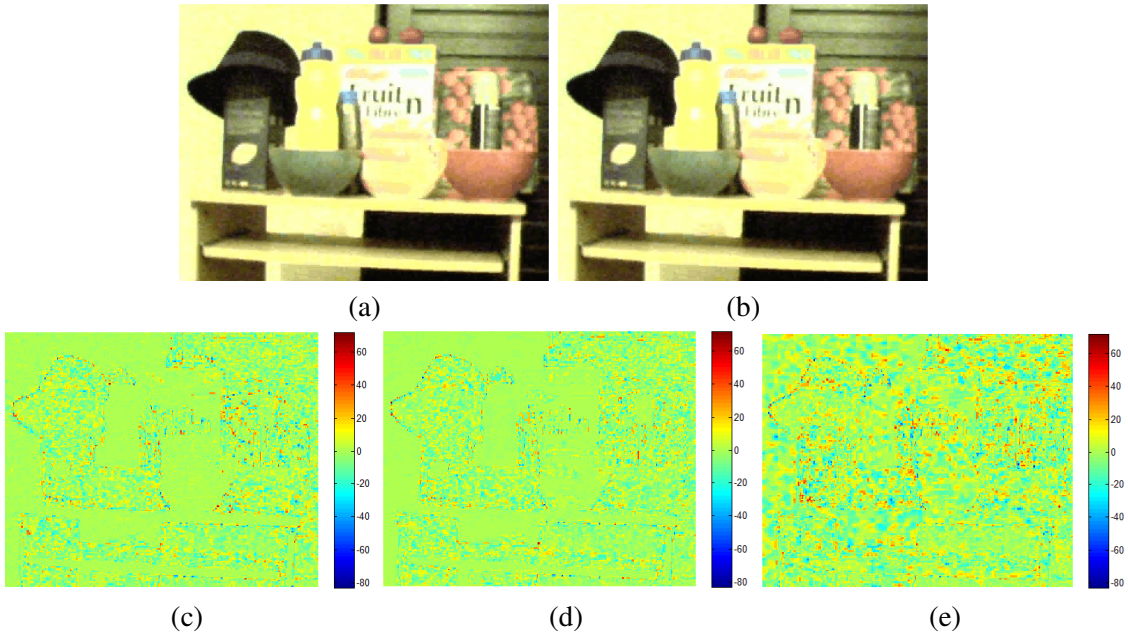


Figure 4.2: Sample consecutive frames, (a) frame 1 and (b) frame 2, of the recorded sequence using a webcamera. The visualisation of the pixel-based difference between the two frames for the corresponding red, green and blue channels is shown in (c), (d) and (e), respectively. The non-zero values in (c), (d) and (e) correspond to the noise in the red, green and blue channels, respectively.

this trial by gradually reducing ζ. On P_8, a set of n_8 test sequences are generated by reducing the resolution of the original video frames by $\rho\%$. Figure 4.3 shows examples of the input generated

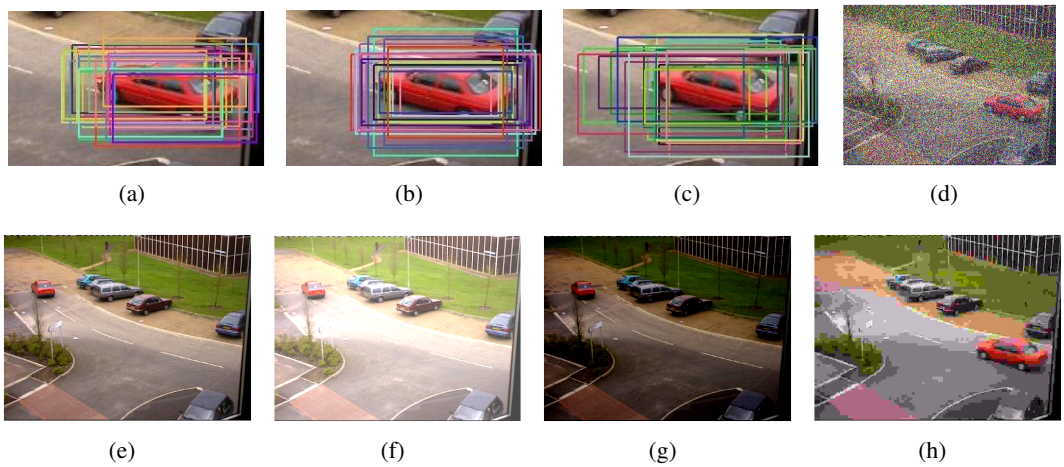


Figure 4.3: Examples of the input generated on different trials. Original images taken from PETS 2000 dataset. (a-c) Visualisation of perturbed initialisations generated on (a) Trial 1, (b) Trial 2 and (c) Trial 3; (d) Trial 4: Addition of zero-mean Gaussian noise with $\sigma^2 = (\sigma_1^2, \sigma_2^2, \sigma_3^2)$ to a test frame; (e-g) Trial 6: (e) original frame and its visualisation after increasing illumination (f) and decreasing illumination (g); (h) Trial 7: visualisation of a test frame after applying compression with $\zeta = 0$.

Table 4.1: Description of eight trials covering various real-world challenges as illustrated in Fig. 4.1. The protocol generates 60 initialisations by adding perturbations to the original (ground-truth) target initialisation and 24 test sequences by modifying the original video.

Real-world challenge	Trial	Description	Parameters
Initialisation errors	P_1	Position	$n_1 = 20$, $\hat{O} = 50$
	P_2	Size	$n_2 = 20$, $\hat{O} = 50$
	P_3	Position and size	$n_3 = 20$, $\hat{O} = 50$
Sensor noise	P_4	Noisy video	$n_4 = 6$, $\hat{l} = 1,2,\ldots,6$
Frame dropping	P_5	Skipping of video frames	$n_5 = 4$, $\hat{m} = 2,4,6,8$
Illumination	P_6	Changing illumination	$n_6 = 2$, $L = 200$
Bandwidth reduction	P_7	Video compression	$n_7 = 4$, $\zeta = 75,50,25,0$
	P_8	Resolution reduction	$n_8 = 8$, $\rho = 10,20,\ldots,80$

on different trials.

Table 4.1 summarises the trials and the values of the corresponding parameters (these parameters accomplished statistically significant results, as discussed at the end of Sec. 4.5). Using the proposed protocol, a tracker is tested on each sequence of the dataset in original form and in its 24 variations generated on different trials. Each tracker is tested with 60 perturbations of the initialisation on the original video sequence. A deterministic tracker is therefore run 85 times, whereas a probabilistic tracker is run $85 \times \mathfrak{n}$ times ($\mathfrak{n} = 10$) for its evaluation using the protocol, where $\mathfrak{n}$ denotes the number of runs for each test of a trial.

4.4 Combined tracking performance score

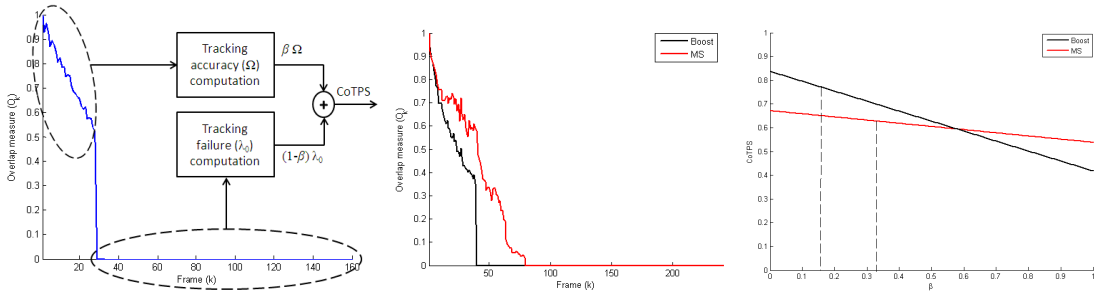
The proposed overlap-based measure effectively summarises single-target tracking performance while taking into account target size variations. The measure is threshold independent and separately quantifies performance in the portion of the sequence where the tracker is successful and where it fails, and combines them into a single score. We describe the proposed measure below.

Without loss of generality, let the estimated target region, A_k, and the ground-truth target region, $\bar{A}_k$, be bounding boxes that define the width and the height of the target. Hence, the estimated trajectory, $\mathfrak{X}$, and the ground-truth trajectory, $\bar{\mathfrak{X}}$, (as defined in Sec. 1.2, Ch. 1) can be re-written as follows:

$$\mathfrak{X} = \{(x_k, y_k, w_k, q_k)\}_{k=k_{ini}}^{k_{end}}, \tag{4.1}$$

$$\bar{\mathfrak{X}} = \{(\bar{x}_k, \bar{y}_k, \bar{w}_k, \bar{q}_k)\}_{k=\bar{k}_{ini}}^{\bar{k}_{end}}, \tag{4.2}$$

where (x_k, y_k) and $(\bar{x}_k, \bar{y}_k)$ define the centroid of the estimated and the ground-truth bounding



(a) (b) (c)

Figure 4.4: (a) The proposed evaluation measure is formulated by combining contributions that quantify tracking accuracy and tracking failure. (b) The result of Mean-Shift (MS) [25] (in red) and of the online boosting-based tracker (Boost) [39] (in black) on the AVSS 2007 sequence containing target h_4 (described in Sec. 4.3). The CoTPS values are 0.628 for MS and 0.770 for Boost. β for MS and Boost is computed using Eq. 4.7. (c) Comparison between values of β computed adaptively. CoTPS plotted for a range of β values for the tracking results of Boost and MS from the example in (b). The interpretation of the performance can significantly change depending on the value of β. A preset value could lead to an incorrect evaluation.

boxes, respectively; and (w_k, q_k) and $(\bar{w}_k, \bar{q}_k)$ represent widths and heights of the estimated and the ground-truth bounding boxes, respectively.

We first compute the amount of overlap, O_k across $\mathfrak{X}$ using Eq. 2.3. Note that $O_k = 0$ if the tracker does not produce a bounding box when the target is present or if a bounding box is produced when no target is present.

The *tracking accuracy* quantifies the extent to which the estimated trajectory overlaps the ground-truth trajectory, considering only frames with $O_k \neq 0$ (Fig. 4.4(a)) and is computed as [62]:

$$\hat{\lambda}^{\hat{\tau}} = \frac{\hat{N}^{\hat{\tau}}}{\hat{N}}, \tag{4.3}$$

where $\hat{N}^{\hat{\tau}} = |\hat{F}^{\hat{\tau}}|$ and $\hat{F}^{\hat{\tau}} = \{f_k : O_k \in (0, \hat{\tau}),\ \hat{\tau} \in (0,1],\ \forall k\}$; and $\hat{N} = |\hat{F}|$, with $\hat{F} = \{f_k : O_k \neq 0,\ \forall k\}$, is the number of frames with $O_k \neq 0$.

Computing $\hat{\lambda}^{\hat{\tau}}$ for a fixed value of $\hat{\tau}$ necessitates an application-dependent decision, since different values of $\hat{\tau}$ may be appropriate for different tracking tasks. To overcome this limitation, instead of computing $\hat{\lambda}^{\hat{\tau}}$ for a fixed value of $\hat{\tau}$, we accumulate its value over the full range of $\hat{\tau}$ values. In particular, we use an increment of $\Delta\hat{\tau} = 0.01$ to obtain $\hat{\lambda}(\hat{\tau})$ and therefore, the score that quantifies tracking accuracy across the sequence, Ω, is computed as

$$\Omega = \Delta\hat{\tau} \sum_{\hat{\tau} \in (0,1]} \hat{\lambda}(\hat{\tau}), \tag{4.4}$$

where $\Omega \in [0,1]$. The smaller Ω, the higher the tracking accuracy. Ω can be regarded as an approximation of the area under the curve of $\hat{\lambda}(\hat{\tau})$.

Tracking failures correspond to instances of target loss. The tracking failure score, λ_0 ($\lambda_0 \in [0,1]$), is defined as

$$\lambda_0 = \frac{N^0}{K}, \tag{4.5}$$

where $N^0 = |F^0|$ is the number of frames with $O_k = 0$ where $F^0 = \{f_k : O_k = 0,\ \forall k\}$ and $K = |F|$ is the total number of frames where $F = \{f_k :\ \forall k\}$. The smaller λ_0, the smaller the tracking failure score.

We combine the information on tracking accuracy and tracking failure in a single score to facilitate performance ranking. The proposed *Combined Tracking Performance Score*, CoTPS (CoTPS $\in [0,1]$), is computed as follows:

$$\text{CoTPS} = \beta\Omega + (1-\beta)\lambda_0, \tag{4.6}$$

where β is a *penalty*, with $\beta \in [0,1]$. The smaller CoTPS, the better the tracking performance. Figure 4.4(b) plots O_k for two tracking results whose comparison is shown using CoTPS.

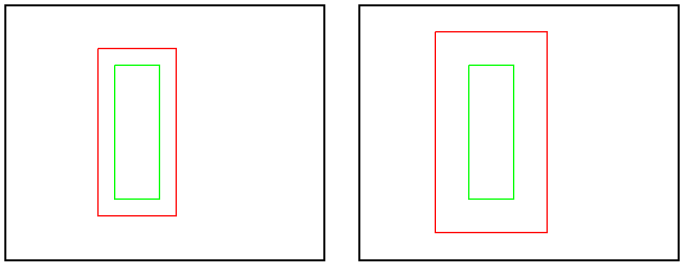
Note that a preset value of β may lead to incorrect performance evaluation (see Fig. 4.4(c)). β is computed adaptively:

$$\beta = \frac{\hat{N}}{K}, \tag{4.7}$$

where $\hat{N}$ is the number of frames in which the tracker has partially or completely tracked the target ($O_k > 0$), thus restricting the inclusion of any extra influence of Ω in the computation of CoTPS. Similarly, $(1-\beta)$ applied to λ_0 is proportional to $(K-\hat{N})$, i.e. the number of frames in which the tracker has failed ($O_k = 0$), which are also the same frames used in the estimation of λ_0, thus restricting the inclusion of any extra influence of λ_0 in the computation of CoTPS.

Let us consider the result of the Mean-Shift tracker (MS) [25] in Fig. 4.4(b). In this example, a penalty of $\beta = 0.328$ (computed using Eq. (4.7)) is applied to Ω since the tracker is successful ($O_k > 0$) in 32.8% frames ($\hat{N} = 79$ and $K = 241$). Similarly, a penalty of $(1-\beta) = 0.672$ is applied to λ_0 since the tracker has failed ($O_k = 0$) in 67.2% frames ($K - \hat{N} = 162$ and $K = 241$). The adaptive computation of β allows us to include accurate contributions of Ω and λ_0 in the estimation of CoTPS.

To conclude this section we show the advantages of CoTPS over existing measures: *Object*

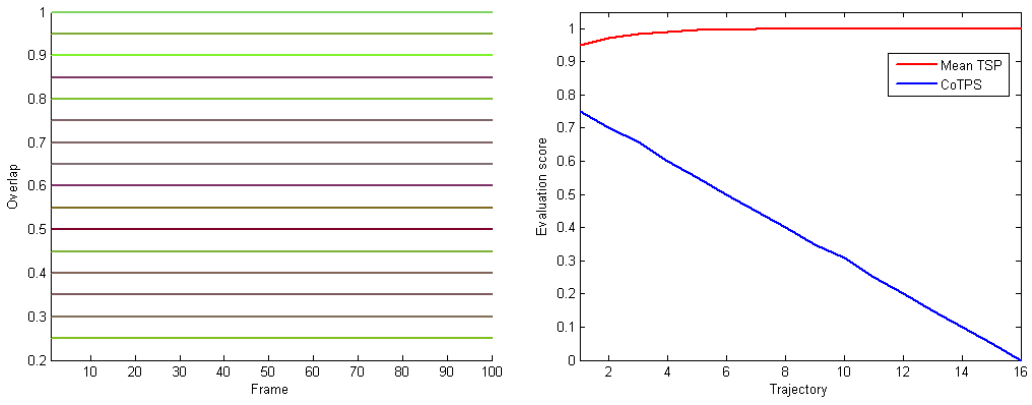


$d_k = 0$ $d_k = 0$
CoTPS=0.550 CoTPS=0.740
(a) (b)

Figure 4.5: Examples showing the advantages of CoTPS over the Object positional accuracy (d_k) [72, 96, 67]. Ground truth: green; tracking results: red. Unlike the Object positional accuracy, CoTPS can distinguish the two different tracking results in (a) and (b).

positional accuracy (d_k) [72, 96, 67] (Eq. 2.1) and the *Tracking Success Probability* (TSP) [60] (Eq. 2.5). Fig. 4.5 shows two examples of tracking with $O_k > 0$ in Fig. 4.5(a) and (b). The Object positional accuracy can not distinguish the two results with $O_k > 0$ and produces for both $d_k = 0$, whereas CoTPS can differentiate them (CoTPS=0.550 and CoTPS=0.740, respectively).

Figure 4.6 compares CoTPS and TSP for a set of 16 trajectories each having a constant overlap with the ground truth for the whole sequence. As in the original paper [60], we use $\nu = 11.8$ and consider the mean TSP score of trajectories in the evaluation (standard deviation of TSP values is equal to zero for all trajectories because of the constant overlap). Unlike CoTPS, the mean TSP scores do not distinguish among the different trajectories (particularly among those having an overlap > 0.45) because of the need of defining 'successful tracking' (i.e. setting the ν parameter) that restricts the variation of TSP scores between 0.95 to 1 for a significant overlap



(a) (b)

Figure 4.6: Comparison between CoTPS and TSP [60] using a set of toy trajectories with constant overlap with the ground truth (a). CoTPS can discriminate between these different types of trajectories, whereas the mean TSP scores cannot (b).

Table 4.2: Description of the dataset. $\grave{C}^h_{ini}, \grave{C}^h_{min}, \grave{C}^h_{max}$ and K denote, in pixels, the initial target size, the minimum target size, maximum target size and the number of frames in the sequence, respectively.

Target	Class	$\grave{C}^h_{ini}$	$\grave{C}^h_{min}$	$\grave{C}^h_{max}$	K	Frame size	Challenges
h_1	Head	139×91	7488	15965	430	576×720	Pose changes, small scale changes
h_2	Head	62×66	370	40128	550	240×320	Pose changes, scale changes, partial occlusions
h_3	Vehicle	227×108	2067	24516	160	576×768	Pose changes, scale changes
h_4	Vehicle	99×103	870	10197	241	576×720	Clutter, pose changes, scale changes
h_5	Person	30×87	180	3444	150	576×768	Partial/total occlusions, pose changes, scale changes
h_6	Person	73×28	638	4410	750	288×384	Partial occlusions, pose changes, scale changes, clutter

range of 0.25 to 1.

4.5 Experimental analysis and validation

We demonstrate the effectiveness of the proposed protocol and score by evaluating and comparing eight state-of-the-art trackers on publicly-available datasets. First we describe datasets and trackers (Sec. 4.5.1). We then present the performance comparison of trackers on each trial, $P_{\hat{i}}$, (Sec. 4.5.2) and with each target, h_t, (Sec. 4.5.3). Finally, we present a discussion and verify the statistical significance of the obtained results. (Sec. 4.5.4).

4.5.1 Dataset and trackers

We selected the dataset by taking into account the diversity of targets and test scenarios, their availability and the challenges involved. The dataset contains three target classes, namely *head*, *vehicle* and *person*. The sequences are chosen from the well-known PETS, CAVIAR, AVSS and SPEVI datasets. A range of tracking challenges are present in the dataset such as partial/total occlusions, pose changes, background clutter and small/large scale changes. The selected sequences include two *head* targets h_1 and h_2 from SPEVI[2], two *vehicle* targets h_3 and h_4 from PETS 2000[3] and AVSS 2007[4], respectively, and two *person* targets h_5 and h_6 from PETS 2010[5] and CAVIAR[6], respectively. Table 4.2 summarises the dataset in terms of initial target size ($\grave{C}^h_{ini}$), minimum and maximum size of the visible part of target ($\grave{C}^h_{min}$ and $\grave{C}^h_{max}$), number of frames (K), frame size and the challenges present in the sequence.

We selected the well-known state-of-the-art trackers including the online boosting tracker (Boost) [39], the semi-supervised online boosting tracker (SemiBoost) [40], the beyond semi-

[2]http://www.eecs.qmul.ac.uk/~andrea/spevi.html. Accessed June 2012.
[3]ftp://ftp.cs.rdg.ac.uk/pub/PETS2000/. Accessed June 2012.
[4]http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html. Accessed June 2012.
[5]http://www.cvg.rdg.ac.uk/PETS2009/a.html#s2l1. Accessed May 2014.
[6]http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/. Accessed June 2012.

Table 4.3: Description of trackers. Key. MS: mean-shift; LmedS: Least Median of Squares-based; n_{sel}: number of selectors, n_{weak}: number of weak classifiers in a selector, W_{search}: search window size, $W_{current}$: current window size; $\grave{R}_{pos}$: search radius for positive samples; $\mathcal{L}_{rate}$: learning rate; n_{ele}: maximum no. of non-zero elements in every row of random matrix.

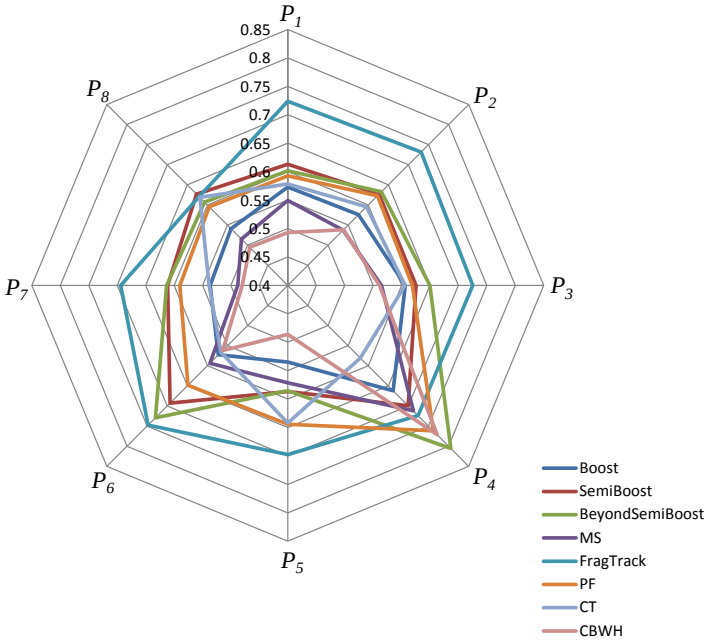
Tracker	Ref.	Description	Parameters
Boost	[39]	Supervised Online boosting	$n_{sel} = 100$, $n_{weak} = 1000$, $W_{search} = 4 \times W_{current}$
SemiBoost	[40]	Semi-supervised online boosting	$n_{sel} = 100$, $n_{weak} = 1000$, $W_{search} = 4 \times W_{current}$
BeyondSemiBoost	[98]	Multi-classifier approach	$n_{sel} = 100$, $n_{weak} = 1000$, $W_{search} = 4 \times W_{current}$
MS	[25]	Histogram matching (RGB), MS minimisation	No. of bins = 4 (each channel), Epanechnikov kernel
FragTrack	[2]	Patch-based histogram matching (gray-level), LMedS minimisation	No. of bins = 16, search window radius = 7 pixels
PF	[78]	Histogram matching (RGB), Monte Carlo framework	No. of bins = 4, No. of particles = 1000
CT	[118]	Compressed feature vectors, Bayes classification	$\grave{R}_{pos} = 4$, $W_{search} = 20$, $\mathcal{L}_{rate} = 0.85$, $n_{ele} = 4$
CBWH	[75]	Corrected background-weighted histogram representation (RGB), MS minimisation	No. of bins=16 (each channel), Epanechnikov kernel

supervised boost (BeyondSemiBoost) [98], the mean-shift tracker (MS) [25], the fragments-based tracker (FragTrack) [2], the particle filter-based tracker (PF) [78], the compressive tracker (CT) [118] and the corrected background-weighted histogram based mean-shift tracker (CBWH) [75]. Each tracker is tested on a total of 187144 frames ($\approx$ 3.25 hours of video data). Tab. 4.3 presents a summary of trackers and lists their parameter values. The parameters of all trackers are fixed throughout the experiments.

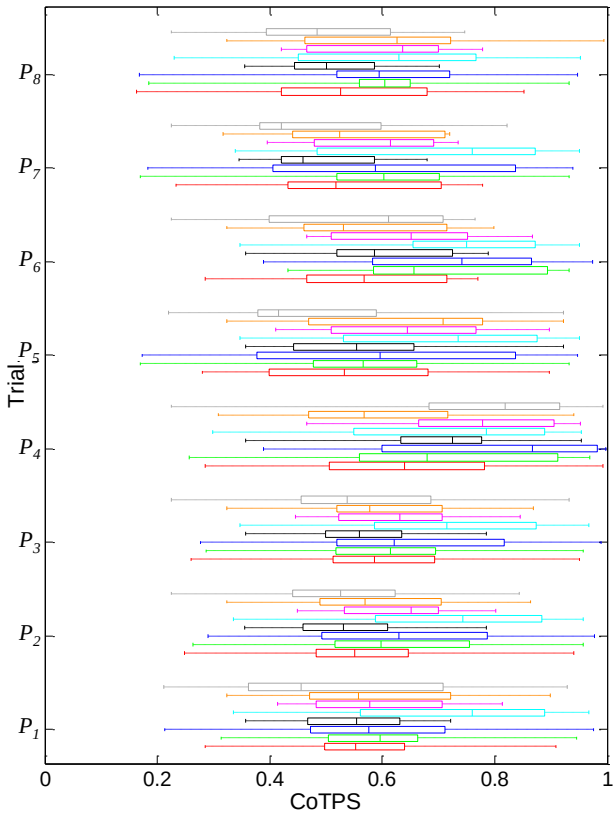
4.5.2 Trial-wise comparison

Fig. 4.7 shows the mean CoTPS (μ_C) of trackers on each P_i computed with all targets and their robustness in terms of the dispersion of their CoTPS (d_C) computed with all targets as $d_C = \text{CoTPS}_{max} - \text{CoTPS}_{min}$, where CoTPS_{max} and CoTPS_{min} are the maximum and the minimum values of CoTPS of a tracker on a trial, respectively.

CBWH and MS consistently track more accurately in the presence of initialisation errors than other trackers (smaller μ_C on P_1, P_2, P_3 in Fig. 4.7(a)). CBWH and MS have a comparable performance on P_2 and P_3, whereas on P_1 CBWH has a better μ_C than MS. The reason for a better performance by CBWH is its improved ability to minimise the background interference in localising a target as compared to MS, enabling the former to handle the initialisation perturbations better [75]. FragTrack shows inferior performance to other trackers in the presence of initialisation errors. In fact, FragTrack uses a fragment-based representation of the target [2] and a perturbation in its initialisation can lead to the inclusion of non-target patches in the target model thus resulting in the accumulation of tracking errors over time. Among the remaining trackers, Boost shows a closer performance to CBWH and MS on these trials (Fig. 4.7(a)) followed by CT, PF and the other two boosting-based trackers. Additionally, in terms of robustness to initialisation errors, MS and PF outperform the remaining trackers (smaller d_C of MS and PF in Fig. 4.7(b)). The reason of the increased sensitivity of the boosting-based trackers (Boost,



(a)



(b)

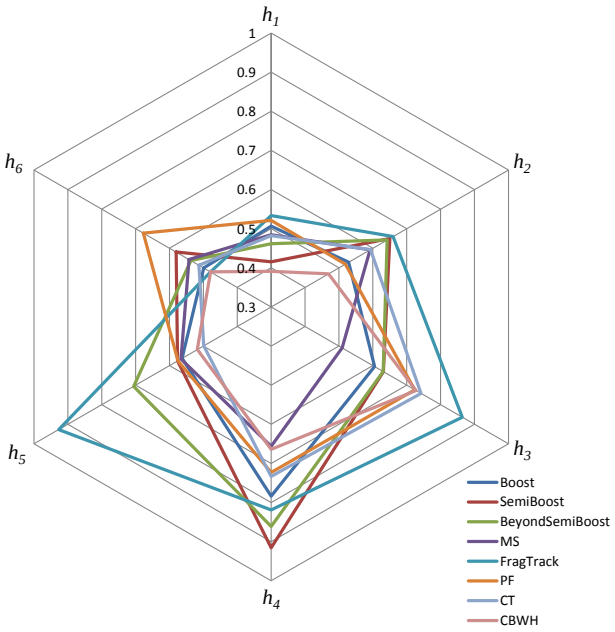
Figure 4.7: Performance comparison of trackers on each trial. (a) Mean CoTPS (μ_C) of trackers on each trial ($P_1, P_2, ..., P_8$) with all targets. (b) CoTPS of trackers on each trial computed with all targets; Boost (red), SemiBoost (green), BeyondSemiBoost (blue), MS (black), FragTrack (cyan), PF (magenta), CT (orange) and CBWH (gray). The dispersion value (d_C) for a tracker is computed as the difference between its maximum and minimum CoTPS values on a trial.

SemiBoost, BeyondSemiBoost) is that any perturbation to initialisation may affect their learning process. The performance of BeyondSemiBoost decreased the most with noisy video data (highest μ_C on P_4). CT and Boost show a better capability at coping with noisy video data due to the online adaptation model (smallest μ_C). PF is more robust at dealing with noise (smaller d_C than the remaining trackers). On P_5, CBWH shows the best performance (smallest μ_C) followed by Boost, MS, BeyondSemiBoost, SemiBoost, CT, PF and FragTrack, respectively. Frame dropping may result in abrupt movements of target: the boosting-based trackers and CBWH are less robust to increasing levels of frame dropping than the remaining trackers. PF is the most robust tracker (smallest d_C on P_5) and BeyondSemiBoost is the least robust. CBWH has the best performance under the changing illumination conditions (smallest μ_C on P_6), showing its ability to adapt to appearance changes. The performance of CT and Boost is closer to CBWH. PF is the most robust when dealing with the changing illumination (smallest d_C on P_6) with the d_C of MS closer to that of PF. An interesting observation regarding the performance of boosting-based trackers on P_6 is that both μ_C and d_C deteriorate from Boost to SemiBoost and from SemiBoost to BeyondSemiBoost, which suggests that the evolution of the boosting-based trackers has resulted in a decreased ability to cope with the changing illumination conditions. The results also highlight the sensitivity of FragTrack to changing illumination (the highest μ_C and the highest d_C on P_6). CBWH and MS have the best performance on P_7 in terms of μ_C. In terms of μ_C, Boost and CT show a closer performance to CBWH and MS, followed by PF, SemiBoost, BeyondSemiBoost and FragTrack, respectively. In terms of d_C, MS and PF have the best performance showing their robustness in coping with the compressed video data, followed by CT, Boost, CBWH, FragTrack, BeyondSemiBoost and SemiBoost, respectively. Finally, on P_8, CBWH and MS are again the best trackers in terms of μ_C. The performance of Boost is closer to CBWH and MS in terms of μ_C as compared to the remaining trackers. Moreover, MS is the most robust in dealing with resolution changes (smallest d_C) with PF showing the closest d_C to MS.

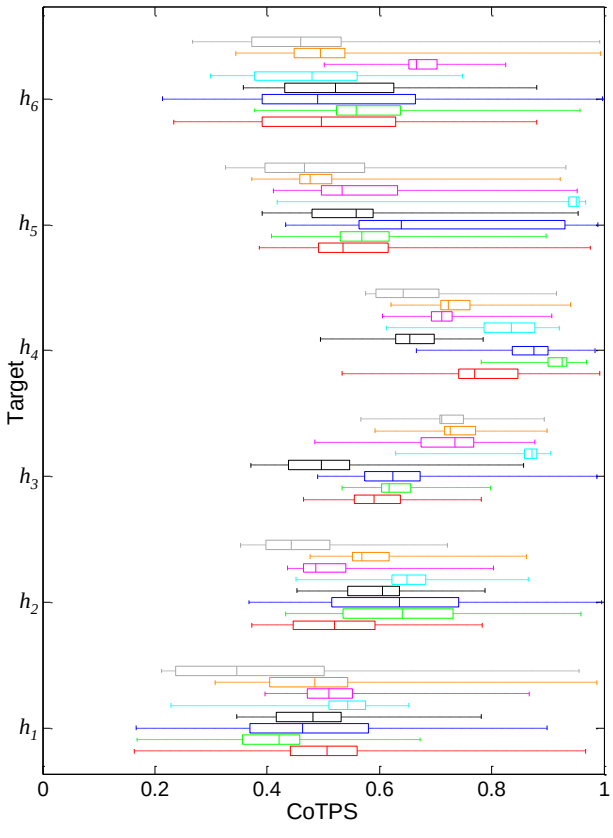
4.5.3 Target-wise comparison

Fig. 4.8 shows the mean CoTPS of trackers (μ_C) on each target ($h_1, h_2, ..., h_6$) and their robustness in terms of dispersion of their CoTPS (d_C) computed in all trials.

h_1 presents the challenges of small scale changes and pose changes. The performance of CBWH is the best on h_1 in terms of μ_C, followed by SemiBoost, BeyondSemiBoost, CT, MS, Boost, PF and FragTrack (Fig. 4.8(a)). In terms of d_C, the results show a smaller variation in

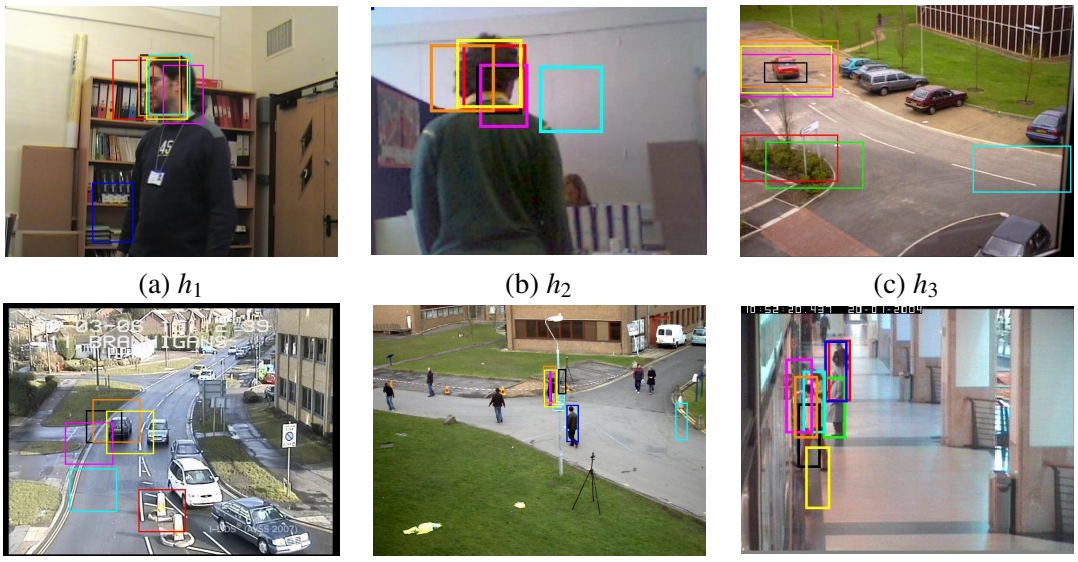


(a)



(b)

Figure 4.8: Performance comparison of trackers on each target. (a) Mean CoTPS (μ_C) of trackers on each target ($h_1, h_2, ..., h_6$) with all trials. (b) CoTPS of trackers on each target computed with all trials; Boost (red), SemiBoost (green), BeyondSemiBoost (blue), MS (black), FragTrack (cyan), PF (magenta), CT (orange) and CBWH (gray). The dispersion value (d_C) for a tracker is computed as difference between its maximum and minimum CoTPS values on a target.



(a) h_1 (b) h_2 (c) h_3

(d) h_4 (e) h_5 (f) h_6

Figure 4.9: Sample tracking results generated by the trackers for the targets, $h_1, \ldots, h_6$. Boost: red; SemiBoost: green; BeyondSemiBoost: blue; MS: black; FragTrack: cyan; PF: magenta; CT: orange; CBWH: yellow.

the performance of FragTrack, MS and PF compared to the remaining trackers (Fig. 4.8(b)). There is a pose change of the target (h_1) around frame 107 of the sequence (Fig. 4.9(a)), where the boosting-based trackers lose the target (Boost only tracks a very small part of target). h_2 presents challenges such as partial occlusions, pose changes and scale changes. CBWH has the best performance in terms of μ_C. The μ_C of PF and Boost is the closest to CBWH. Moreover, MS has a smaller variation (d_C) in its performance on h_2 compared to the remaining trackers. A significant pose change of target (360° turning) at around frame 145 (Fig. 4.9(b)) causes SemiBoost, BeyondSemiBoost and FragTrack to lose the target. h_3 undergoes gradual change in its scale and pose across the sequence, which are handled by MS that tracks consistently well followed by PF, CT and CBWH (Fig. 4.9(c)). Indeed, MS outperforms the other trackers in terms of μ_C. This is because the appearance of target h_3 is bright and well-distinguished from the background, and the use of colour distribution enables MS to track well on the various generated test sequences containing h_3. SemiBoost has the smallest variation in performance on h_3 (smallest d_C). h_4 is challenging due to the presence of background clutter, similar objects (vehicles) and scale changes, and all trackers obtain a higher μ_C. MS and CBWH have the best performance on h_4 in terms of μ_C followed by PF, CT, Boost, FragTrack, BeyondSemiBoost and SemiBoost, respectively. In terms of variation in performance, although d_C for SemiBoost is the smallest, this is less important as its CoTPS is mostly very high. The appearance of h_4 is quite similar to

that of the road, making it challenging to track. All trackers have generally struggled to track this target with CT, CBWH and MS showing a better tracking (Fig. 4.9(d)). h_5 presents challenges including occlusions, scale changes and pose changes. CT has the best performance in terms of μ_C (μ_C of CBWH is closer to CT). SemiBoost shows a smaller variation (d_C) in its performance on h_5 as compared to other trackers. h_5 faces a severe occlusion around frame 51 where only CT, CBWH and PF can track the target after the occlusion (Fig. 4.9(e)). h_6 has challenges such as partial occlusions, small pose changes and clutter. FragTrack and CBWH outperform the other trackers in terms of μ_C. The sequences containing h_2 and h_5 also involve pose changes and partial occlusions but FragTrack has not performed as well on them (Fig. 4.8(a)). In fact, h_2 involves significant pose changes and h_5 involves severe occlusions, suggesting that FragTrack can cope better with small pose changes and partial occlusions [2]. Fig. 4.9(f) shows frame 359 involving partial occlusion where FragTrack performs well (CT and PF also track a small part of the target). Finally, in terms of d_C PF shows the best performance.

4.5.4 Discussion

Fig. 4.10 shows the performance of trackers for each target class (*head, vehicle, person*). All trackers except CT have their best performance on *head* followed by *person* and *vehicle*. CT shows the best performance on *person* followed by *head* and *vehicle*. The overall best performance on *head* and *person* tracking is by CBWH. The performance of CT is closer to CBWH on *person* tracking. The overall best performance on *vehicle* tracking is by MS. There is an inconsistency in the performance of FragTrack on *person* tracking: while it has achieved the best performance on h_6, its performance reduces significantly on h_5 (Fig. 4.8(a)), as discussed earlier.

Fig. 4.11 shows the *cumulative performance* of trackers on all trials and all targets. CBWH has the best perfomance in terms of μ_C followed by MS, Boost, CT, PF, SemiBoost, Beyond-SemiBoost and FragTrack, respectively. PF is more robust than the remaining trackers as shown by its smaller d_C. Finally, overall, the boosting-based trackers are less robust (higher d_C in Fig. 4.11) when dealing with various test scenarios than the remaining trackers. CBWH and MS more capably handle initialisation errors and outperform other trackers with compressed videos and resolution reductions. CBWH also copes better with the frame dropping and changing illumination conditions than the rest of the trackers. CT shows the best performance in the presence of noisy video data. Moreover, among the boosting-based trackers, Boost handles pose changes

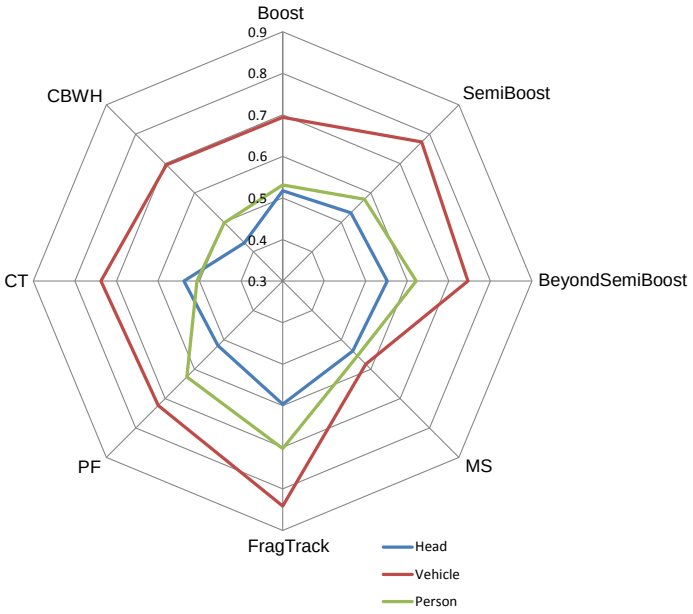


Figure 4.10: Performance comparison of trackers on each target class (*head, vehicle, person*) based on mean CoTPS.

better than SemiBoost and BeyondSemiBoost. Among the remaining trackers, MS, PF, CT and CBWH can handle small as well as large pose changes; whereas FragTrack can only deal with small pose changes. PF and CT deal with partial and total occlusions better than all the other trackers.

Finally, we tested the statistical significance of CoTPS using the Welch ANOVA test [109], a modified version of the one-way ANalysis Of VAriance (ANOVA) test [34] commonly employed to test statistical significance of multiple groups of data (in our case, there are eight groups each containing a set of CoTPS of a tracker) whose variances are unequal [68]. Statistical significance was achieved on each trial, on each target and on each target class at the standard significance level $\alpha = 0.05$.

4.6 Summary

In this chapter, we presented a new evaluation protocol that enables testing the robustness of trackers under various real-world conditions. These are encapsulated in a series of trials. We also introduced a new overlap-based criterion for quantifying the performance of single-target trackers on trials. The proposed criterion evaluates performance by combining tracking accuracy and tracking failure scores. An extensive experimental analysis and validation is presented in the form of a statistically significant performance comparison of eight state-of-the-art trackers. The

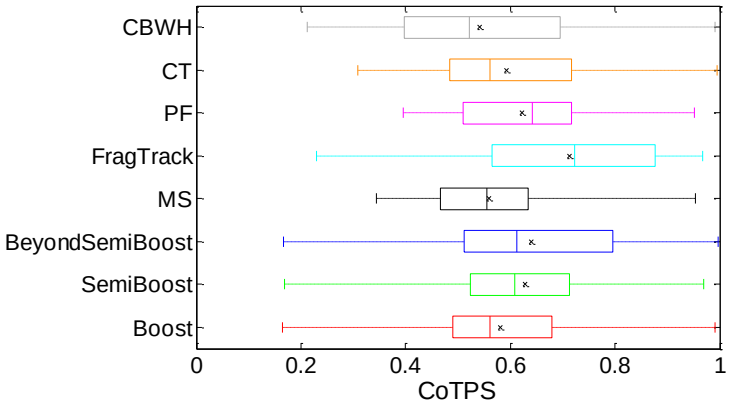


Figure 4.11: Cumulative performance of trackers: the mean CoTPS of trackers computed on all trials with all targets are shown with a '$\times$' in the corresponding boxplots. The dispersion value (d_C) for a tracker is computed as a difference between maximum and minimum CoTPS values in its boxplot.

implementation of the protocol is available online (http://www.eecs.qmul.ac.uk/~andrea/pft2) to provide the research community with a platform to present and compare the performance of their trackers.

While the effectiveness of the proposed evaluation protocol is shown in combination with the proposed measure (CoTPS), the trials are indeed generic and can also be used with other state-of-the-art measures. It is therefore relevant to analyse and compare different measures in order to understand their strengths and weaknesses. To this end, in the next chapter we shall propose a methodology to *quantitatively* assess the relative performance of different measures.

Chapter 5

Assessment of tracking evaluation measures

5.1 Introduction

Rapid advances in video tracking research [2, 25, 40, 98, 118] have been accompanied by new proposed performance evaluation measures [10, 60, 89, C3, J2], which in turn need to be systematically assessed in order to understand their relative performances. Discrepancy-based empirical measures evaluate performance by quantifying the deviation of tracking results from a ground truth over time at frame level [60] or at sequence level [89]. The measures may evaluate tracking performance based for example on the extent of spatial match between the tracked region and the ground-truth target region over time. The spatial match may be determined in the form of the amount of common pixels [60] or coincidence between the tracked and ground-truth regions [10]. Coincidence is defined as the existence of the centroid of one region within the other region. In the tracking evaluation domain, the focus is mainly laid on the performance comparison of trackers [77, 115], whereas the quantitative assessment of evaluation measures is missing. We present a methodology for the quantitative assessment of discrepancy-based evaluation measures with respect to human judgement [C2]. The comparison and analysis are based on determining the probabilistic agreement between the decisions made by measures and made by humans on tracking results (Fig. 5.1).

In this chapter, we first define the problem for the assessment of measures in Sec. 5.2. Sec. 5.3 describes the evaluation measures to be assessed. Sec. 5.4 describes the subjective evaluation procedure with respect to which the measures will be assessed in Sec. 5.5. The chapter

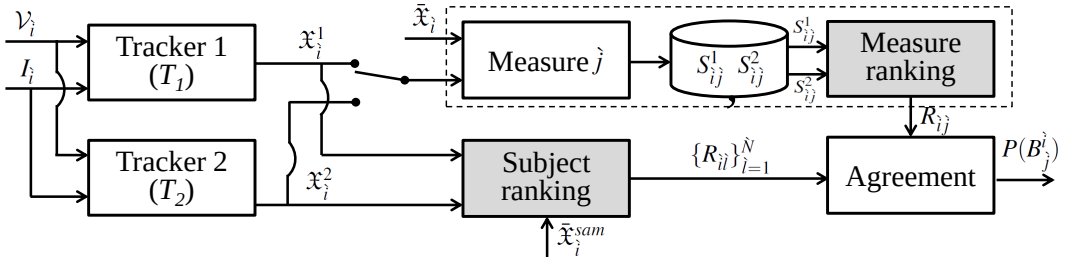


Figure 5.1: Empirical assessment of measures with respect to human judgement. T_1 and T_2 tested on video clip ($\mathcal{V}_{\grave{i}}$) with initialisation ($I_{\grave{i}}$). $S^1_{\grave{i}\grave{j}}$ and $S^2_{\grave{i}\grave{j}}$ are the performance scores computed using the measure $\grave{j}$ by evaluating $\mathfrak{X}^1_{\grave{i}}$ and $\mathfrak{X}^2_{\grave{i}}$, the estimated trajectories of T_1 and T_2 on $\mathcal{V}_{\grave{i}}$, respectively, with respect to ground-truth trajectory $\bar{\mathfrak{X}}_{\grave{i}}$. $R_{\grave{i}\grave{j}}$ is the decision of the measure $\grave{j}$ based on $S^1_{\grave{i}\grave{j}}$ and $S^2_{\grave{i}\grave{j}}$. $R_{\grave{i}\grave{l}}$ is the decision of the human subject $\grave{l} : \grave{l} = 1,\ldots,\grave{N}$, based on $\mathfrak{X}^1_{\grave{i}}$ and $\mathfrak{X}^2_{\grave{i}}$ while using also the shown ground-truth samples, $\bar{\mathfrak{X}}^{sam}_{\grave{i}}$. $P(B^{\grave{j}}_{\grave{i}})$ denotes the amount of agreement on $\mathcal{V}_{\grave{i}}$ between $R_{\grave{i}\grave{j}}$ and the set of human judgements, $\{R_{\grave{i}\grave{l}}\}^{\grave{N}}_{\grave{l}=1}$.

is summarised in Sec. 5.6.

5.2 Problem definition

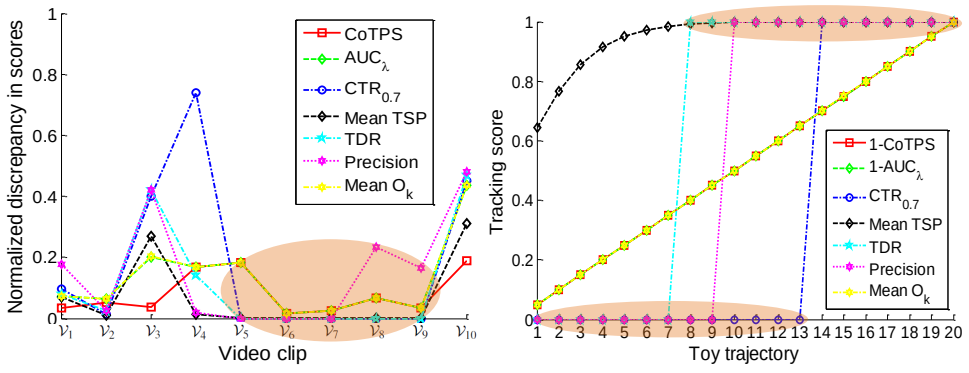
Let us consider two trackers, T_1 and T_2, run on $\grave{M}$ video clips, $\mathcal{V}_{\grave{i}} : \grave{i} = 1,...,\grave{M}$, to track a target with initialisation, $I_{\grave{i}}$. We use a pair of trackers to ease the subjective comparison made by humans. T_1 and T_2 generate the trajectories $\mathfrak{X}^1_{\grave{i}}$ and $\mathfrak{X}^2_{\grave{i}}$, respectively, in each video clip ($\mathcal{V}_{\grave{i}}$). $\mathfrak{X}^1_{\grave{i}}$ and $\mathfrak{X}^2_{\grave{i}}$ are a sequence of states over frames: $\mathfrak{X}^1_{\grave{i}}=\{X^1_{\grave{i}k}\}^{K^1_{\grave{i}}}_{k=1}$, where $X^1_{\grave{i}k}$ is the estimated state of T_1 at frame k of $\mathcal{V}_{\grave{i}}$, and $K^1_{\grave{i}}$ is the number of frames where $\mathfrak{X}^1_{\grave{i}}$ exists. $X^1_{\grave{i}k}$ may contain information about the target position $(x^1_{\grave{i}k}, y^1_{\grave{i}k})$ and the occupied region $A^1_{\grave{i}k}$: $X^1_{\grave{i}k}=\{(x^1_{\grave{i}k}, y^1_{\grave{i}k}), A^1_{\grave{i}k}\}$. Let $\bar{\mathfrak{X}}_{\grave{i}}$, $\bar{X}_{\grave{i}k}$, $\bar{K}_{\grave{i}}$, $(\bar{x}_{\grave{i}k}, \bar{y}_{\grave{i}k})$ and $\bar{A}_{\grave{i}k}$ represent the corresponding ground-truth of the quantities defined above. $\mathfrak{X}^1_{\grave{i}}$ and $\mathfrak{X}^2_{\grave{i}}$ are evaluated with respect to $\bar{\mathfrak{X}}_{\grave{i}}$ using one out of $\grave{J}$ measures ($\grave{j} = 1,...,\grave{J}$) to obtain their corresponding evaluation scores, $S^1_{\grave{i}\grave{j}}$ and $S^2_{\grave{i}\grave{j}}$, respectively.

Based on the comparison between $S^1_{\grave{i}\grave{j}}$ and $S^2_{\grave{i}\grave{j}}$ we define the rank $R_{\grave{i}\grave{j}}$ as follows: $R_{\grave{i}\grave{j}}=(1,2)$ if $S^1_{\grave{i}\grave{j}}$ is better than $S^2_{\grave{i}\grave{j}}$; $R_{\grave{i}\grave{j}}=(2,1)$ if $S^2_{\grave{i}\grave{j}}$ is better than $S^1_{\grave{i}\grave{j}}$; or $R_{\grave{i}\grave{j}}=(1.5,1.5)$ if $S^1_{\grave{i}\grave{j}}=S^2_{\grave{i}\grave{j}}$. $R_{\grave{i}\grave{j}}=(1.5,1.5)$ defines a tie between T_1 and T_2 [48]. Similarly, let $R_{\grave{i}\grave{l}}$ be the judgement (decision) of the $\grave{l}$th human subject (*s.t.* $\grave{l} = 1,...,\grave{N}$) in ranking $\mathfrak{X}^1_{\grave{i}}$ and $\mathfrak{X}^2_{\grave{i}}$. $R_{\grave{i}\grave{l}}$ is defined as $R_{\grave{i}\grave{j}}$, where $\grave{j}$ in $R_{\grave{i}\grave{j}}$ is substituted with $\grave{l}$. We aim to assess the measure $\grave{j}$ by quantifying the amount of agreement between $R_{\grave{i}\grave{j}}$ and the judgements of subjects, $\{R_{\grave{i}\grave{l}}\}^{\grave{N}}_{\grave{l}=1}$ (Fig. 5.1).

5.3 Measures

This section lists the evaluation measures to be assessed with respect to human judgements. We consider the following representative state-of-the-art single-target tracking evaluation measures: Tracking Success Probability (TSP) [60] (Ch. 2, Eq. 2.5), Mean Dice (MD) vs Correct Track Ratio (CTR) curve [89] (Ch. 2), Track Detection Rate (TDR) [10] (Ch. 2, Eq. 2.17), Precision ($\hat{P}$) (Ch. 2, Eq. 2.7), Mean Overlap ($\overline{O}$) [54], Area Under lost-track ratio Curve (AUC_λ) [C3] (Ch. 2, Eq. 2.10) and Combined Tracking Performance Score (CoTPS) [J2] (Ch. 4, Eq. 4.6). TSP, MD-vs-CTR curve and $\hat{P}$ involve presetting of parameters in the evaluation procedure, whereas TDR, AUC_λ, CoTPS and $\overline{O}$ do not require presetting of parameters. In the case of TSP, we use the mean TSP score ($\overline{\text{TSP}}$) across the trajectory and the fixed parameter ν=11.8 [60]. In the case of MD-vs-CTR curve, to quantify the tracking performance we use the CTR value corresponding to MD of at least 0.7, i.e. $\min\{\text{MD}\}_{MD\geq 0.7}$, denoted as $CTR_{0.7}$. $CTR_{0.7} \in [0,1]$: the higher $CTR_{0.7}$, the better the result. A Dice score ≥ 0.7 is considered to be a satisfactory tracking result [89]; hence the threshold of 0.7 is used for $CTR_{0.7}$, thus showing the long-term tracking ability as the percentage of the sequence where the target is tracked with MD of at least 70%. In the case of $\hat{P}$, we use $\tau_2 = 0.25$ for head targets and $\tau_2 = 0.50$ for person and vehicle targets as done in [8]. $\overline{O}$ is computed as the average of the overlap, O_k, (Ch. 2, Eq. 2.3) across the trajectory where the target exists.

Here we analyse a property of measures that is the ability to distinguish different tracking results. Such a property is important when evaluating in applications where even a very slight tracking discrepancy needs to be quantified. Fig. 5.2(a) shows the normalised discrepancy between evaluation scores of each measure for tracker pairs on $\grave{M}$ video clips (where $\grave{M} = 10$ as discussed in the next section), which is the absolute difference between the evaluation scores of tracker pairs computed using the measure divided by its range i.e. (upper bound - lower bound). $\overline{O}$, AUC_λ and CoTPS consistently distinguish tracker pairs on all clips (normalised discrepancy > 0), whereas the remaining measures are unable to distinguish results (i.e. normalised discrepancy = 0) from $\mathcal{V}_5$ to $\mathcal{V}_9$ as highlighted in Fig. 5.2(a), except $\hat{P}$ that could distinguish performance on $\mathcal{V}_8$ and $\mathcal{V}_9$. To further analyse the distinguishing ability of the measures, we consider a set of 20 toy trajectories each having a constant overlap (for the whole sequence) of $0.05, 0.10, \ldots, 1$, respectively. The overlap is as $a(\cdot)$ (defined in Eq. 2.6) for $\overline{\text{TSP}}$, as O_k (defined in Eq. 2.3) for AUC_λ, CoTPS and $\hat{P}$, and as D_k (defined in Eq. 2.4) for $CTR_{0.7}$. For TDR, coincidence is achieved throughout a



(a) (b)

Figure 5.2: Ability of the measures to distinguish different tracking results. (a) Normalised discrepancy in the evaluation scores of each measure for the tracking pair on each clip, $\mathcal{V}_1, \ldots, \mathcal{V}_{10}$. (b) Tracking scores are computed using different measures for 20 toy trajectories.

trajectory when $O_k \geq 0.4$ (i.e. for trajectory 8 to trajectory 20). We plot the scores computed by measures for 20 trajectories in Fig. 5.2(b). We can clearly see two groups of measures. The first group includes (1-CoTPS), (1-AUC_λ) and $\overline{O}$, which can each discriminate the results throughout overlap variations. The second group includes $\overline{\text{TSP}}$, $\hat{P}$, $CTR_{0.7}$ and TDR, which often are not able to distinguish variations in results (as highlighted in Fig. 5.2(b)) due to their preset thresholds on the overlap or coincidence.

Next we describe the subjective evaluation performed to gather decisions of human subjects (on video clips) with respect to which the measures will be quantitatively assessed.

5.4 Subjective evaluation

We use ten test video clips ($\mathcal{V}_1$ to $\mathcal{V}_{10}$) with different target types (head, vehicle, person), challenges (scale change, pose change, occlusion, clutter) and scenarios (indoor, outdoor). The video clips are from publicly-available datasets including AVSS 2007 challenge[1], CAVIAR[2], Clemson head tracking[3], PETS 2000[4], PETS 2010[5] and SPEVI[6] (Tab. 5.1, Fig. 5.3). Additionally, we use the state-of-the-art trackers including the mean-shift tracker (MS) [25], the particle filter-based tracker (PF) [78], the fragments-based tracker (FragTrack) [2], the online boosting tracker (Boost) [39], the semi-supervised online boosting tracker (SemiBoost) [40] and the beyond semi-

[1]http://www.eecs.qmul.ac.uk/~andrea/avss2007_d.html. Accessed June 2012.
[2]http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/. Accessed June 2012.
[3]http://www.ces.clemson.edu/~stb/research/headtracker/seq/. Accessed February 2014.
[4]ftp://ftp.cs.rdg.ac.uk/pub/PETS2000/. Accessed June 2012.
[5]http://www.cvg.rdg.ac.uk/PETS2009/a.html#s2l1. Accessed May 2014.
[6]http://www.eecs.qmul.ac.uk/~andrea/spevi.html. Accessed June 2012.

Table 5.1: Summary of the dataset. Frame size is in pixels. Key. K: number of frames in $\mathcal{V}_{\grave{i}}$.

Video clip	$\mathcal{V}_1$	$\mathcal{V}_2$	$\mathcal{V}_3$	$\mathcal{V}_4$	$\mathcal{V}_5$	$\mathcal{V}_6$	$\mathcal{V}_7$	$\mathcal{V}_8$	$\mathcal{V}_9$	$\mathcal{V}_{10}$
Dataset	Clemson head tracking				SPEVI		PETS2000	AVSS2007	PETS2010	CAVIAR
K	51	83	50	50	100	29	30	30	30	100
Duration (sec)	7	11	6	7	7	4	4	4	4	11
Frame size	96×128	96 × 128	96 × 128	96 × 128	576 × 720	240 × 320	576 × 768	576 × 720	576 × 768	288 × 384
Target	Head						Vehicle		Person	

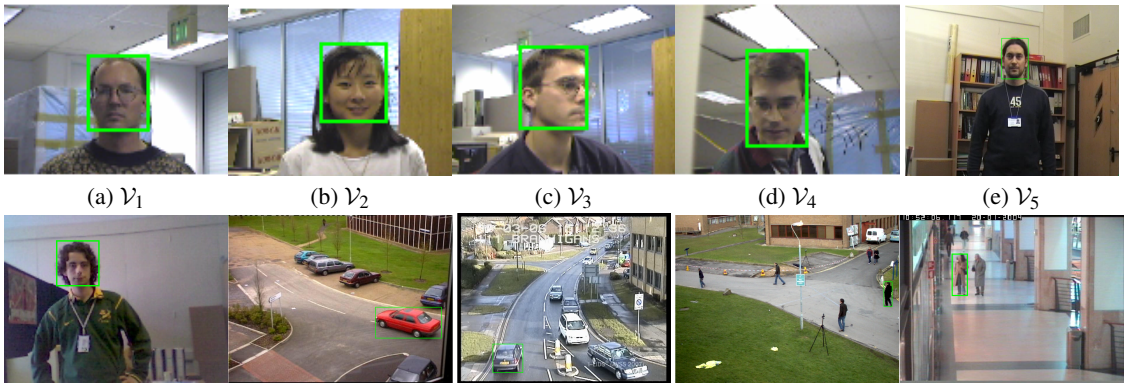
supervised boost (BeyondSemiBoost) [98].

We asked human subjects to rank the pair of trackers' results $(\mathfrak{X}_{\grave{i}}^1, \mathfrak{X}_{\grave{i}}^2)$ on all $\mathcal{V}_{\grave{i}}$. For each $\mathcal{V}_{\grave{i}}$, the trackers' results are provided to subjects in qualitative form with $\mathfrak{X}_{\grave{i}}^1$, $\mathfrak{X}_{\grave{i}}^2$ superimposed as a sequence of bounding boxes over time. Three samples of subjects are distinguished as *skilled*, *semi-skilled* and *unskilled* in target tracking, which is determined by the subjects *per se*. $\grave{N}_1$, $\grave{N}_2$ and $\grave{N}_3$ denote the size of the skilled, semi-skilled and unskilled samples (in our study $\grave{N}_1 = \grave{N}_2 = \grave{N}_3 = 30$). As suggested in [61], we selected subjects that were not involved in our work.

The subjective evaluation tests were done on a website[7] that, after providing the instructions, shows the tracking results of tracker pairs (T_1, T_2) side by side. The gray colour of the background (red=green=blue=130) of the webpage follows the recommendation by ITU for relaxing human eyes [49]. For each clip, the ground-truth tracking samples are also provided as reference for the first, middle and last frame. We show short clips (4-11 sec) to help human subjects remember the tracking results, thereby minimising the uncertainty in their judgment. The clips are played in a loop and can be viewed multiple times. Each subject chooses the tracker, "Left" or "Right", which they deem to be the best or chooses "Same" if the result of each tracker in the pair appears indistinguishable.

We aim to statistically analyse the decisions of skilled, semi-skilled and unskilled subjects on all video clips. To this end we test the statistical significance for decisions of a sample of judges (i.e. skilled or semi-skilled or unskilled subjects). We define two hypotheses, the *null hypothesis* (H_0) and *alternate hypothesis* (H_a), as follows. H_0: a set of judges cannot distinguish the performance of two trackers on a video; H_a: a set of judges can distinguish the performance of two trackers on a video. We statistically check whether H_a is valid by rejecting H_0 according to a level of significance, α. The level of significance indicates the probability of rejecting a true null hypothesis. α is often set to 0.05 [48]. We intend to choose a test that can be applied for ranked data and account for ties. We choose the Friedman's Two-Way ANOVA test (Friedman's

[7] http://webprojects.eecs.qmul.ac.uk/fabiop/subjeval/. Accessed January 2014.



(a) $\mathcal{V}_1$ (b) $\mathcal{V}_2$ (c) $\mathcal{V}_3$ (d) $\mathcal{V}_4$ (e) $\mathcal{V}_5$

(f) $\mathcal{V}_6$ (g) $\mathcal{V}_7$ (h) $\mathcal{V}_8$ (i) $\mathcal{V}_9$ (j) $\mathcal{V}_{10}$

Figure 5.3: Visualisation of the first frame of video clips used for subjective evaluation. The targets are indicated with green bounding boxes. Datasets: (a-d) Clemson head tracking, (e-f) SPEVI, (g) PETS 2000, (h) AVSS 2007 challenge, (i) PETS 2010 and (j) CAVIAR.

test) [48]. Friedman's test, χ^2, is computed as

$$\chi_r^2 = \frac{12}{\grave{N}\grave{F}(\grave{F}+1)} \sum_{\grave{f}=1}^{\grave{F}} \left(\sum_{\grave{l}=1}^{\grave{N}} R_{\grave{i}\grave{l}}(\grave{f}) \right)^2 - 3\grave{N}(\grave{F}+1), \tag{5.1}$$

where $R_{\grave{i}\grave{l}}(\grave{f})$ is the rank assigned to tracker $T_{\grave{f}}$ on $\mathcal{V}_{\grave{i}}$ by subject $\grave{l}$, such that $\grave{f}$=$\{1,2\}$ because we consider a pair of trackers ($\grave{F}$=2). To test the statistical significance at α=0.05, the χ^2 value is compared to the value corresponding to ($\grave{F}$-1) degrees of freedom in the χ^2 table of critical values [48] that is equal to 3.841. If $\chi^2 > 3.841$, the statistical significance is achieved and H_0 is rejected.

We therefore perform the Friedman's test on each $\mathcal{V}_{\grave{i}}$ for skilled ($\grave{N}$=$\grave{N}_1$), semi-skilled ($\grave{N}$=$\grave{N}_2$) and unskilled ($\grave{N}$=$\grave{N}_3$) samples separately. The statistical significance is achieved for all $\mathcal{V}_{\grave{i}}$ except for $\mathcal{V}_6$ where all the three categories did not achieve it and for $\mathcal{V}_2$ where semi-skilled subjects did not achieve it (Fig. 5.4). The reason for not achieving the statistical significance on $\mathcal{V}_6$ is that the results seem very comparable (Fig. 5.5(e-g), Fig. 5.6(f)) due to which the three categories of subjects could not distinguish tracking results (Fig. 5.7(a)); hence a very small χ_r^2 (Fig. 5.4). Moreover, the reason for not achieving statistical significance on $\mathcal{V}_2$ for semi-skilled subjects is their discordant judgements: 53.33% subjects ranked T_2 to be better than T_1; 26.67% ranked T_1 to be better than T_2; and 20.00% considered T_1 and T_2 to be the same. Indeed, the results of T_1 and T_2 are close to each other across the sequence (Fig. 5.6(b), Fig. 5.5(a-c)), until T_1 completely loses the target in the last frame (Fig. 5.6(b), Fig. 5.5(d)). Unskilled subjects have less discordant decisions with $\chi_r^2 = 4.80$ (Fig. 5.4): 63.33% subjects ranked T_2 to be better than T_1; 23.33% ranked T_1 to be better than T_2; and 13.33% considered T_1 and T_2 to be the same. As for skilled subjects, 80.00% ranked T_2 to be better than T_1; 13.33% ranked T_1 to be better than

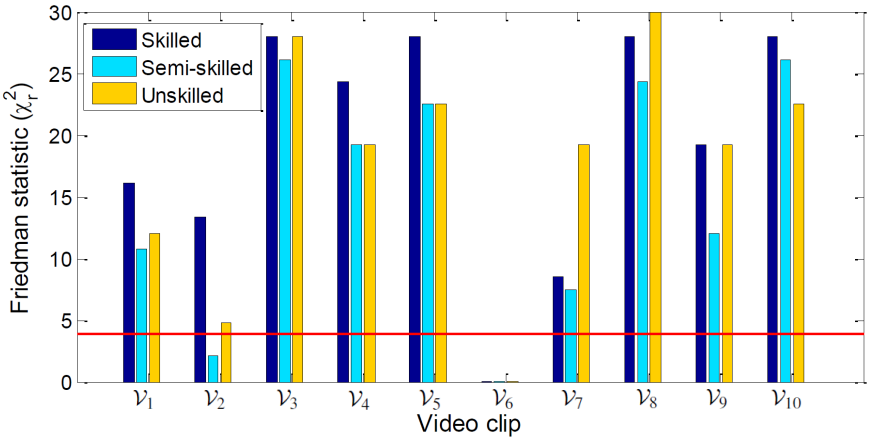


Figure 5.4: Statistical significance using Friedman test (χ^2) on each $\mathcal{V}_i$ for skilled, semi-skilled and unskilled (subject) samples. The red line indicates the critical value corresponding to the standard significance level, $\alpha = 0.05$.

T_2; and 6.67% considered T_1 and T_2 to be the same. Hence, a much higher $\chi_r^2 = 13.33$.

5.5 Assessment of measures

We devise a probabilistic criterion to assess measures that is based on computing their agreement with the judgements of subjects (measure-subject agreement). Let us consider a set of events for a sample of subjects (skilled, semi-skilled or unskilled) in a probability space for each $\mathcal{V}_i$, which is defined as follows: $\mathbf{E}^i=\{E_1^i, E_2^i, E_3^i\}$: E_1^i= {T_1 is better than T_2 on $\mathcal{V}_i$}; E_2^i= {T_2 is better than

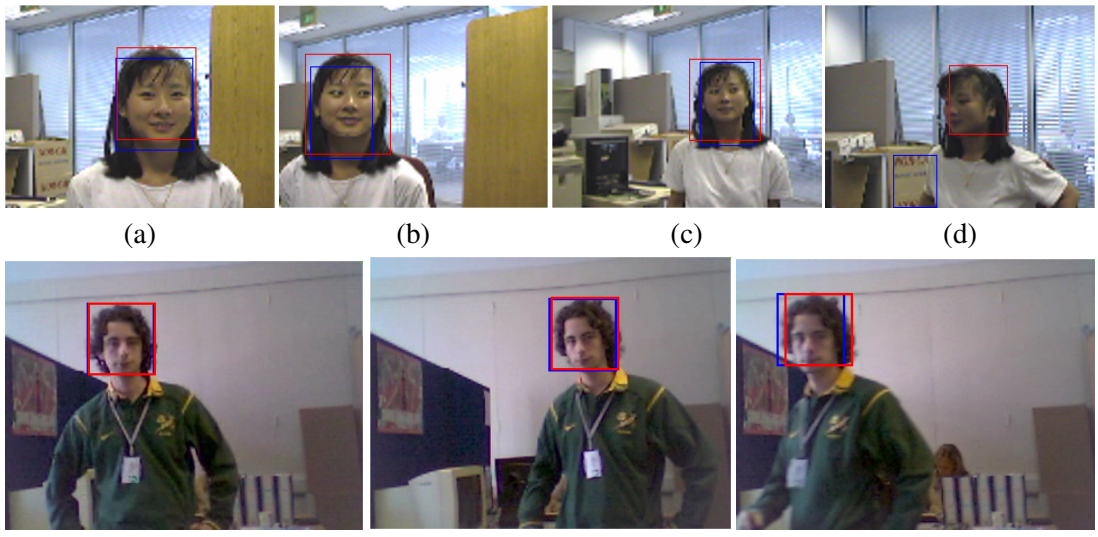
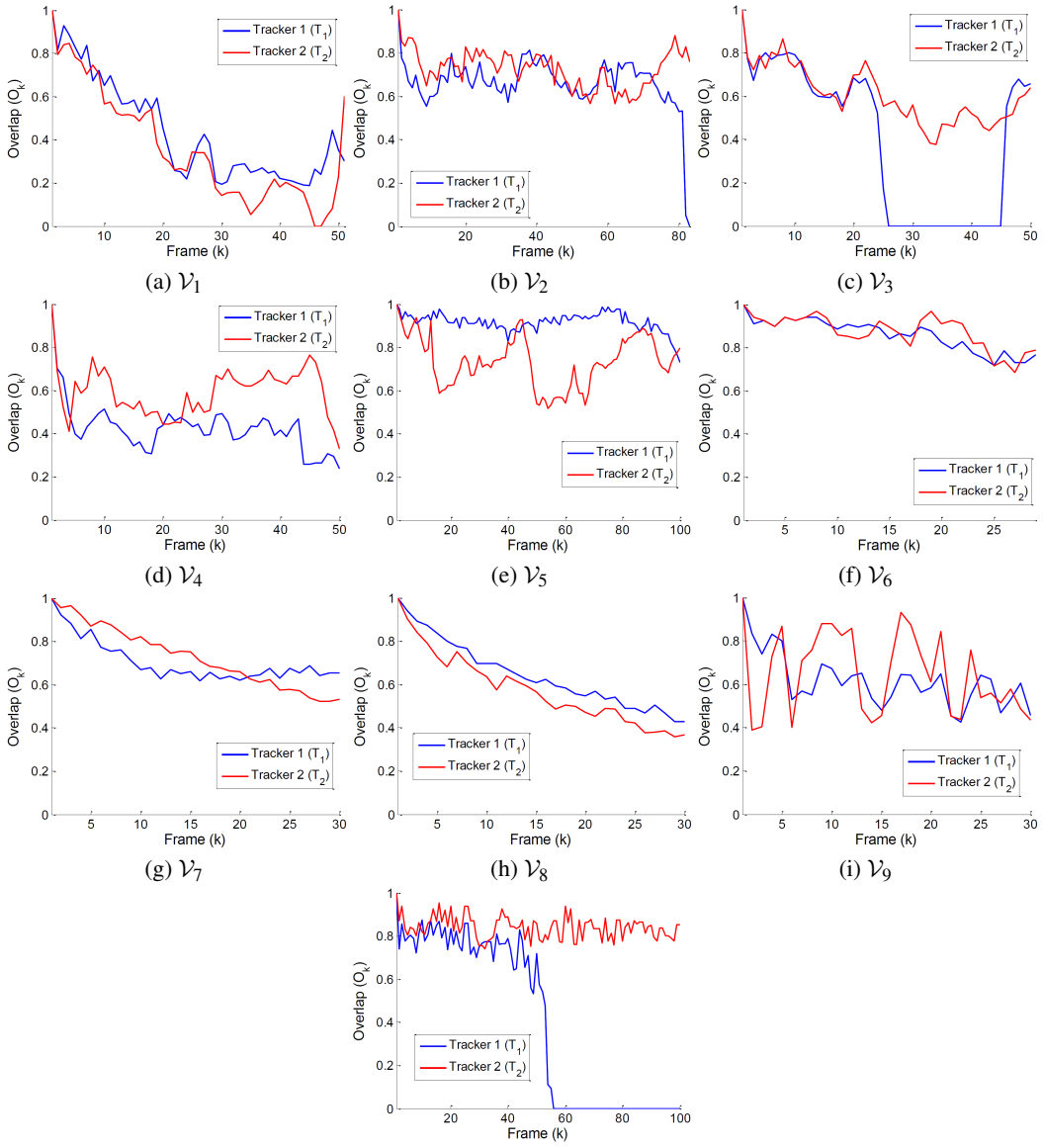


Figure 5.5: Examples of result of T_1 (blue) and T_2 (red): (a-d) frame 3, 32, 65, 83 of $\mathcal{V}_2$; (e-g) frames 5, 15, 24 of $\mathcal{V}_6$.



(a) $\mathcal{V}_1$ (b) $\mathcal{V}_2$ (c) $\mathcal{V}_3$

(d) $\mathcal{V}_4$ (e) $\mathcal{V}_5$ (f) $\mathcal{V}_6$

(g) $\mathcal{V}_7$ (h) $\mathcal{V}_8$ (i) $\mathcal{V}_9$

(j) $\mathcal{V}_{10}$

Figure 5.6: Amount of overlap (common pixels), O_k, between estimated and ground-truth results for pair of trackers (T_1, T_2).

T_1 on $\mathcal{V}_{\grave{i}}\}$; $E_3^{\grave{i}}$= $\{T_1$ and T_2 are the same on $\mathcal{V}_{\grave{i}}\}$.

We can compute the probability of an event occurring on $\mathcal{V}_{\grave{i}}$ as

$$P(E_{\grave{q}}^{\grave{i}}) = \frac{n_{E_{\grave{q}}^{\grave{i}}}}{n_{E_1^{\grave{i}}} + n_{E_2^{\grave{i}}} + n_{E_3^{\grave{i}}}}, \quad \forall \grave{q} = 1,2,3, \tag{5.2}$$

where $n_{E_{\grave{q}}^{\grave{i}}}$ denotes the number of times $E_{\grave{q}}^{\grave{i}}$ occurs for a $\mathcal{V}_{\grave{i}}$ and for each sample. We find the probability, $P(B_{\grave{j}})$, of the $\grave{j}$th measure ($B_{\bar{\grave{j}}}^{\grave{i}}$ has the same probability space as $E_{\grave{q}}^{\grave{i}}$) by calculating

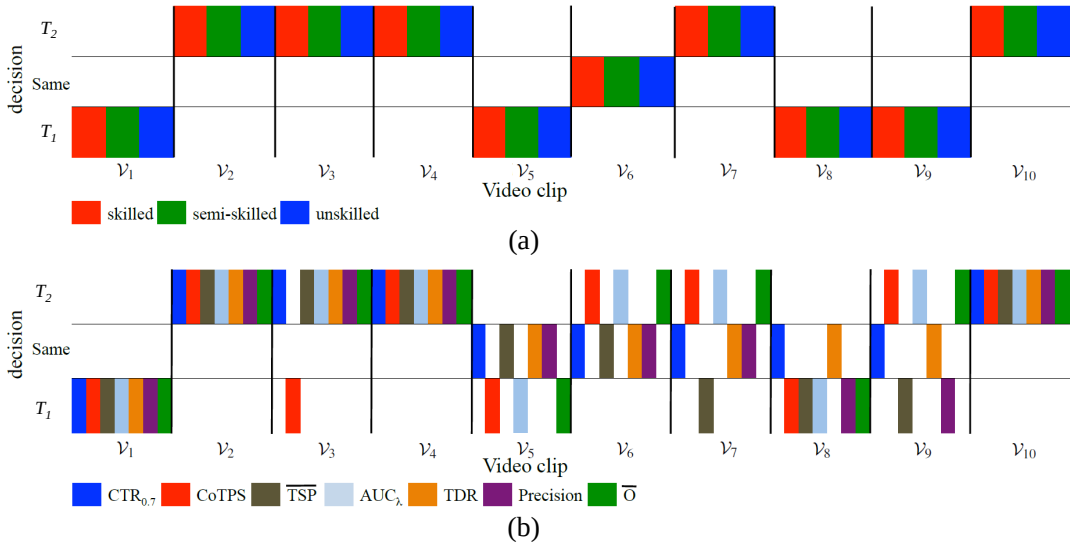


Figure 5.7: Decision for each video sequence ($\mathcal{V}_{\grave{i}}$). The decision regarding ranking between the tracker pair (T_1, T_2) taken on each $\mathcal{V}_{\grave{i}}$ by (a) (most of) the skilled, semi-skilled and unskilled subjects, (b) the evaluation measures. 'T_1', 'T_2' and 'Same' on the vertical axis show T_1 considered the best, T_2 considered the best and both trackers considered the same, respectively.

the total probabilities for $\grave{M}$ independent sets of events computed for each sample of subjects:

$$P(B_{\grave{j}}) = \frac{1}{\grave{M}} \sum_{\grave{i}=1}^{\grave{M}} \sum_{\grave{q}=1}^{3} P(B_{\grave{j}}^{\grave{i}} | E_{\grave{q}}^{\grave{i}}) P(E_{\grave{q}}^{\grave{i}}), \tag{5.3}$$

where $\grave{M}$ is the normalisation factor. We use $P(B_{\grave{j}})$ to quantify the agreement between the $\grave{j}$th measure and each sample of subjects (i.e. skilled, semi-skilled and unskilled) (Tab. 5.2).

The measures with the overall highest agreement with the three subject samples are $\hat{P}$ and $\overline{\text{TSP}}$ (Tab. 5.2). AUC_λ and $\overline{O}$ also consistently achieve high $P(B_{\grave{j}})$. CoTPS is in agreement with AUC_λ and $\overline{O}$ on all $\mathcal{V}_{\grave{i}}$ (Fig. 5.7(b)) except on $\mathcal{V}_3$, which resulted in a lower $P(B_{\grave{j}})$ for CoTPS. $CTR_{0.7}$ and TDR show the lowest $P(B_{\grave{j}})$ for the three subject samples.

CoTPS, AUC_λ and $\overline{O}$ can capture slight changes in tracking results even in the cases when humans show uncertainty in distinguishing them. The ability to capture these changes is useful in accurately ranking the tracking results and quantifying even the minor discrepancy (that may be desirable for some applications). For example, these three measures can distinguish the trackers on $\mathcal{V}_6$ by judging T_2 as better (Fig. 5.7(b)), despite the fact that the majority of skilled (97%), semi-skilled (90%) and unskilled (90%) subjects judge them as indistinguishable. A limitation in CoTPS can be seen on $\mathcal{V}_3$ where T_1 is judged to be better than T_2, which is opposite to the judgement of the remaining measures and subjects as well. This limitation is due to the non-linear (quadratic) behaviour of CoTPS (Eq. 4.6) due to its failure term (since $\lambda_0 = 1 - \beta$). To

Table 5.2: Assessment in terms of the measure agreement ($P(B_{\hat{j}})$) with the skilled, semi-skilled and unskilled subject samples. The brighter the cell, the better (higher) the agreement.

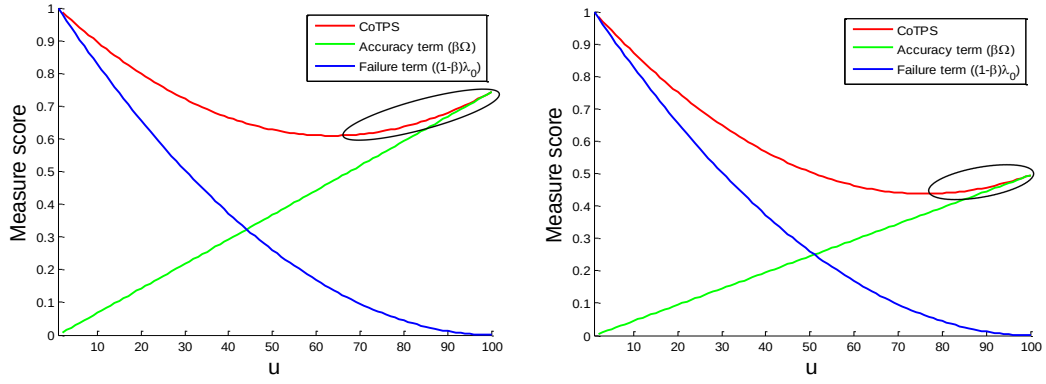
Measure	$\overline{\text{TSP}}$	$\hat{P}$	$CTR_{0.7}$	CoTPS	AUC_{λ}	$\overline{O}$	TDR
Skilled	0.74	0.74	0.58	0.61	0.71	0.71	0.58
Semi-skilled	0.68	0.67	0.52	0.57	0.66	0.66	0.52
Unskilled	0.70	0.71	0.53	0.61	0.70	0.70	0.53

explain it, let us consider a trajectory of length K^{toy} such that there is a constant overlap ($O_k = \hat{c}$: $\hat{c} \in (0,1]$ is a constant) from frame 1 to frame u and tracking failure ($O_k = 0$) from frame u to frame K^{toy}, where $1 \leq u \leq K^{toy}$. While varying u, we compute and plot the corresponding CoTPS, the contributions of its accuracy term ($\beta\Omega$) and its failure term ($(1-\beta)\lambda_0$), when $\hat{c} = 0.25, 0.50$ (Fig. 5.8). For increasing u, CoTPS is expected to be decreasing due to fewer frames having tracking failures. However, this trend may not be visible as encircled in the plots due to its non-linearity (Fig. 5.8). $\overline{\text{TSP}}$ and $\hat{P}$ are mostly in agreement (Fig. 5.7(b)) and also with respect to the subjects (Tab. 5.2). $\overline{\text{TSP}}$ and $\hat{P}$ indeed penalise bad tracking results and poorly discriminate between good results (Fig. 5.2(b). TDR and $CTR_{0.7}$ have the lowest agreement ($P(B_{\hat{j}})$) with subjects and have a limited ability to distinguish tracking results. Fig. 5.7(b) shows that 50% of video clips are judged the 'Same' and this does not correspond to the judgment of subjects (Fig. 5.7(a)). Additionally, the smallest $P(B_{\hat{j}})$ of TDR indicates that tracking evaluation based on the coincidence criterion is not reflecting humans' judgments.

Overall, this study reveals that $\hat{P}$ and $\overline{\text{TSP}}$ are the best measures in terms of their agreement ($P(B_{\hat{j}})$) with human judgement. Due to the hard decisions caused by their preset thresholds on the overlap, $\hat{P}$ and $\overline{\text{TSP}}$ however show a lesser ability to distinguish tracking results. Indeed, $\hat{P}$ and $\overline{\text{TSP}}$ could be used with the proposed trials in Ch. 4 and would be desirable if the trackers are to be grouped into performance classes on each trial. Alternatively, if the goal is to achieve a more distinct (clearly delineated) ranking of trackers, the parameter-independent measures are expected to be suitable due to their better ability to distinguish variations in tracking results. For example, $\overline{O}$ could be chosen in such a case that has an improved distinguishing ability and, at the same time, shows a substantially high $P(B_{\hat{j}})$ as well.

5.6 Summary

In this chapter, we proposed a methodology to empirically assess tracking measures based on the law of total probability that quantifies the agreement between their decisions and those of



(a) $\hat{c} = 0.25$ (b) $\hat{c} = 0.50$

Figure 5.8: A limitation in CoTPS due to the non-linear behaviour.

human subjects in terms of ranking trackers' results. The results unveiled interesting aspects of the assessed measures. $\hat{P}$ and $\overline{\text{TSP}}$ exhibit the best performance because of their highest agreement with humans. These measures, however, have a limited ability to distinguish tracking results. $CTR_{0.7}$ and TDR showed the lowest agreement. AUC_{λ} and $\overline{O}$ are parameter independent, have a better ability to distinguish results and show a substantially high agreement with humans (although lower than $\hat{P}$ and $\overline{\text{TSP}}$).

Chapter 6

Conclusions

6.1 Summary of achievements

This thesis focused on devising ground-truth-based trajectory evaluation measures and procedures to objectively (no use of fixed threshold parameters) quantify tracking performance. We presented a set of measures to account for different aspects of single-target and multi-target tracking evaluation, a protocol to enable the trackers' evaluation under several real-world operational conditions, and a methodology to quantitatively assess the evaluation measures. We elaborate below on the specific achievements of this thesis.

The state-of-the-art multi-target tracking evaluation measures need to pre-set threshold parameters [86, 52, 117], do not take into account target-size variations [86, 10, 44], are not numerically bounded [86, 52], and do not evaluate cardinality error [10, 44]. We proposed three measures [J1] to account for the key aspects of extended multi-target tracking evaluation, which are accuracy, cardinality error and ID changes. The proposed measures do not require parameter presetting, are numerically bounded and take into account the target-size changes over time. The first measure, Multiple Extended-target Tracking Error (METE), uses an effective trade-off between accuracy and cardinality errors to quantify tracking performance at frame level. To separately analyse the accuracy and cardinality error contributions in the computation of METE, we calculate the accuracy error rate and cardinality error rate. The formulation of METE is inspired from OSPA [86] without the need of including the parameters (c and p). OSPA is applicable for the case of point-target representation, whereas METE is applicable for the case of extended-

target representation. The second measure, Multiple Extended-target Lost-Track ratio (MELT), evaluates accuracy at sequence level to highlight the long-term tracking capability. MELT allows tracking performance analysis at varying accuracy levels that can help to choose a tracker for a specific application. The third measure, Normalised ID Changes (NIDC), evaluates ID changes while taking into account the length of the track in which they occur. We performed a detailed experimental validation and analysis to show the effectiveness of the proposed measure using state-of-the-art multi-target trackers and real-world challenging datasets. We showed the advantages of the proposed measures over well-known existing measures. MODA [52] does not have a lower numerical bound, whereas METE is numerically bounded. Additionally, due to presetting of the overlap threshold (τ_o) MODA can not distinguish track pairs (of estimated and ground-truth results) for an overlap variation of $O_k \in [0, \tau_o)$ and $O_k \in [\tau_o, 1]$, whereas METE does not require threshold presetting and can better distinguish different tracking results. Moreover, MODA quantifies the tracking performance by considering false positives and false negatives without accounting for true positives, whereas METE provides a more thorough performance assessment by implicitly considering also true positives in the evaluation procedure. MOTP [52] does not consider track pairs with an overlap value lower than τ_o that may exclude some of the tracking information in the evaluation, whereas MELT provides a more holistic performance evaluation by taking into account all track pairs thus including all of the tracking information. The use of normalisation in NIDC helps to compare the trackers that generate tracks of different lengths, to assess the trackers ability to track for a long duration without confusing the IDs and to compare trackers' performance across different datasets. We also showed the usefulness of the proposed measures in terms of the evaluation of trackers. The results and analysis show that DP-NMS [81] is the best tracker in terms of accuracy error on all datasets (except TownCentre) but it has reported the highest cardinality error on all datasets (except iLids Easy). CRFBT [116] is the best tracker in terms of maintaining the unique IDs of targets over time, followed by MT-TBD [83]. MCMCDA [8] performs better as a person tracker than as a head tracker. Finally, the software implementation of the measures is provided online (http://www.eecs.qmul.ac.uk/~andrea/mtte.html), which is expected to facilitate the community in order to holistically evaluate and compare their trackers.

An important aspect while evaluating the performance of a tracker is to test its robustness in the presence of different operational conditions that refer to the distortions potentially induced

to tracker's input in a real application. To this end we introduced a protocol [J2, C3] to evaluate the trackers on a range of real-world operational conditions including initialisation errors by a detector, sensor noise, frame dropping possibly caused during video data transmission or due to the delayed generation of results by the tracker, changing illumination in the scene and video compression. These conditions are synthetically created and encapsulated into a series of predefined evaluation procedures called trials. We showed the effectiveness of proposed protocol for single-target trackers and to quantify their performance on trials we also proposed a parameter-independent and numerically-bounded single-target tracking evaluation measure, called Combined Tracking Performance Score (CoTPS) [J2], for extended targets. CoTPS combines the contributions from the tracking accuracy and tracking failure terms to provide a single-score performance evaluation that can facilitate trackers' ranking on trials. We performed an extensive experimentation for the proposed framework by providing the performance analysis and comparison of several state-of-the-art trackers on real datasets. The usefulness of the framework lies in enabling the selection of trackers for specific operational conditions. The main findings are as follows. CBWH [75] and MS [25] are the best at handling initialisation errors, compressed video data and resolution reductions. CBWH shows the best performance at dealing with frame dropping and changing illumination conditions. CT [118] performs the best in the presence of noisy video data. Overall, CBWH is the best tracker based on the evaluation on all the trials. Additionally, based on the analysis we also identified the strengths of trackers in dealing with different challenges in the video sequences. MS, PF [78], CT, CBWH and Boost [39] are more capable in handling target pose changes. PF and CT are the best in coping with occlusions. MS, CT and CBWH are better in dealing with target scale changes than the remaining trackers. The software implementation of the protocol is available online (http://www.eecs.qmul.ac.uk/~andrea/pft2) to facilitate the community to present and compare the performance of their trackers. This work [J2, C3] has been used and cited by [19, 79, 65, 22, 28, 35] and VOT Challenge 2013 [54, 55], and is expected to be beneficial for the community in the future. It is also relevant to mention that the proposed trials can also be used in combination with other measures.

The introduction of a variety of tracking evaluation measures in the literature is not accompanied with a systematic procedure to quantitatively assess their relative performance. Indeed, the choice of measures may bias the outcome of the comparison of tracking results. In this regard, we presented a methodology to quantitatively assess the tracking measures by quantifying the

amount of agreement between their decisions, i.e. the ranking of trackers' results, with respect to those of human subjects [C2]. We generated the reference by performing the subjective evaluation test online to gather the decisions of a set of 90 subjects in terms of ranking the pairs of trackers on ten video clips from real publicly-available datasets. We showed the usefulness of the proposed methodology by assessing seven single-target tracking evaluation measures to provide their relative performance in terms of the extent of their agreement with human judgements. Additionally, we also analysed measures along another dimension, which was to highlight their ability to distinguish different results. The results show that $\hat{P}$ and $\overline{\text{TSP}}$ [60] are the best measures because of their highest agreement with humans. These measures, however, can cause hard decisions and show a lesser ability to distinguish tracking results due to the need to preset threshold parameters. Therefore, they could be used to evaluate trackers' performance on the trials of the proposed protocol when a distinct or clearly delineated ranking of trackers is not required. For obtaining a distinct ranking a threshold-independent measure (e.g. $\overline{O}$ [54]) is expected to be appropriate because of the better ability to capture slight variations in tracking results. Moreover, we identified a limitation in CoTPS due to its non-linear behaviour, which may result in an incorrect performance evaluation of trackers.

6.2 Future work

Below are discussed the future directions of this thesis work:

1. The proposed measures for single-target tracking evaluation in Ch. 4 (CoTPS) and multi-target tracking evaluation in Ch. 3 (METE, MELT, NIDC) are suitable for targets that are modeled in terms of their position and 2D image-plane-occupied area. These measures can therefore also be applicable in the case of other sensing modalities provided the same (2D) target model is considered. Moreover, since other target models for 2.5D and 3D tracking also exist for different sensing modalities [99, 85, 110], future work could focus in applying the idea of the proposed measures to such higher-dimensional models.

2. We showed the usefulness of the proposed protocol in terms of the evaluation of single-target trackers (Ch. 4). The trials in the protocol are indeed generic and can be used for evaluating multi-target trackers using the multi-target tracking evaluation measures. Future work could focus on using the trials in combination with multi-target tracking measures to evaluate the robustness of multi-target trackers. Moreover, the trials are not designed

specifically for a ground-truth-based evaluation only and can therefore also be used in combination with standalone evaluation criteria [91].

3. We showed the effectiveness of the proposed measures' assessment methodology in terms of a quantitative evaluation of single-target tracking measures (Ch. 5). As the methodology is generic, future work could focus in applying it to assess multi-target tracking measures using multi-target trackers' results.

4. We identified a limitation of CoTPS in our analysis in Ch. 5, which is caused due to its non-linear behaviour because of the contribution from its failure term (β in Eq. 4.6). Future work could involve addressing this issue with an improved combination of the failure and accuracy terms in CoTPS (Eq. 4.6).

Bibliography

[1] Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definitions. *International Organization for Standardization*, ISO 5725-1, December 1994.

[2] A. Adam, E. Rivlin, and I. Shimshoni. Robust fragments-based tracking using the integral histogram. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 798–805, USA, 2006.

[3] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(8):1619–1632, August 2011.

[4] V. Badrinarayanan, P. Perez, F. Le Clerc, and L. Oisel. On uncertainties, random features and object tracking. In *Proc. of Int. Conf. on Image Processing*, San Antonio, 2007.

[5] S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, March 2011.

[6] F. Bashir and F. Porikli. Performance evaluation of object detection and tracking systems. In *Proc. of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pages 7–14, 2006.

[7] A. Baumann, M. Boltz, J. Ebling, M. Koenig, H. S. Loos, M. Merkel, W. Niem, J. K. Warzelhan, and J. Yu. A review and comparison of measures for automatic video surveillance systems. *EURASIP Journal on Image and Video Processing*, 2008.

[8] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3457–3464, Colorado Springs, USA, Jun. 2011.

[9] T. A. Biresaw, A. Cavallaro, and C. S Regazzoni. Tracker-level fusion for robust bayesian visual tracking. *IEEE Trans. on Circuits and Systems for Video Technology*, to appear.

[10] J. Black, T. Ellis, and P. Rosin. A novel method for video tracking performance evaluation. In *Proc. of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pages 125–132, 2003.

[11] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(9):1820–1833, Sep. 2011.

[12] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, New York City, June 2006.

[13] L. M. Brown, A. W. Senior, Y.-L. Tian, J. Connell, A. Hampapur, C. f. Shu, H. Merkl, and M. Lu. Performance evaluation of surveillance systems under varying conditions. In *Proc. of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–8, 2005.

[14] A. Buchanan and A. Fitzgibbon. Combining local and global motion models for feature point tracking. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, June 2007.

[15] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2000.

[16] A. Bulling and H. Gellersen. Toward mobile eye-based human-computer interaction. *IEEE Pervasive Computing*, 9(4):8–12, 2010.

[17] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby. Detecting anomalies in people's trajectories using spectral graph analysis. *Computer Vision and Image Understanding*, 115(8):1099–1111, August 2011.

[18] S. Calderara, A. Prati, and R. Cucchiara. Mixtures of von mises distributions for people trajectory shape analysis. *IEEE Trans. on Circuits and Systems for Video Technology*, 21(4):457–471, 2011.

[19] L. Cehovin, M. Kristan, and A. Leonardis. Is my new tracker really better than yours? In *Proc. of IEEE Winter Conference on Applications of Computer Vision*, Steamboat Springs CO, March 2014.

[20] D. Chau, F. Bremond, and M. Thonnat. Online evaluation of tracking algorithm perfor-

mance. In *Proc. of Int. Conf. on Imaging for Crime Detection and Prevention*, London, UK, December 2009.

[21] D. P. Chau, F. Bremond, and M. Thonnat. Online evaluation of tracking algorithm performance. In *Proc. of Int. Conf. on Imaging for Crime Detection and Prevention*, London, 2009.

[22] C.-H. Chen, Y. Yao, A. Koschan, and M. Abidi. A novel performance evaluation paradigm for automated video surveillance systems. *Central European Journal of Computer Science*, 1(4):430–441, December 2011.

[23] R. Collins, X. Zhou, and S.K. Teh. An open source tracking testbed and evaluation web site. In *Proc. of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, pages 17–24, January 2005.

[24] R. T. Collins, Y. Liu, and M. Leordeanu. Online selection of discriminative tracking features. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1631–1643, October 2005.

[25] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(5):564–577, May 2003.

[26] Axis Communications. An explanation of video compression techniques. http://www.axis.com/files/whitepaper/wp_videocompression_33085_en_0809_lo.pdf.

[27] P. Correia and F. Pereira. Standalone object segmentation quality evaluation. *EURASIP Journal on Applied Signal Processing*, 4:389–400, 2002.

[28] M. L. Cruz. Evaluacion de algoritmos de seguimiento de objetos. Master's thesis, Escuela Politecnica Superior, Universidad Autonoma de Madrid, 2012.

[29] A. Doulamis. Dynamic tracking re-adjustment: a method for automatic tracking recovery in complex visual environments. *Multimedia Tools and Applications*, 50(1):49–73, October 2010.

[30] T.J. Ellis. Performance metrics and methods for tracking in surveillance. In *Proc. of Int. Workshop on Visual Surveillance-Performance Evaluation of Tracking and Surveillance*, Copenhagen, June 2002.

[31] C. E. Erdem, B. Sankur, and A.M. Tekalp. Performance measures for video object segmentation and tracking. *IEEE Trans. on Image Processing*, 13(7):937–951, 2004.

[32] C. E. Erdem, A.M. Tekalp, and B. Sankur. Video object tracking with feedback of performance measures. *IEEE Trans. on Circuits and Systems for Video Technology*, 13(4):310–324, April 2003.

[33] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sep. 2010.

[34] R. A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, 98:39–82, 1935.

[35] VA Frantz, VV Voronin, VI Marchuk, AV Fisunov, and MM Pismenskova. An algorithm for constructing the path of movement of objects in a video stream based on optical flow. *Electronic Scientific Journal*, 2013 (http://ivdon.ru/magazine/archive/n3y2013/1856).

[36] B. E. Fridling and O. E. Drummond. Performance evaluation methods for multiple-target-tracking algorithms. In *Proc. SPIE, Signal and Data Processing of Small Targets*, volume 1481, pages 371 – 383, 1991.

[37] J. Garcia, A. Gardel, I. Bravo, J. L. Lazaro, M. Martinez, and D. Rodriguez. Directional people counter based on head tracking. *IEEE Trans. on Industrial Electronics*, 60(9):3991–4000, 2013.

[38] A. Godil, R. Bostelman, K. Saidi, W. Shackleford, G. Cheok, M. Shneier, and T. Hong. 3d ground-truth systems for object/human recognition and tracking. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Portland, Oregon, 2013.

[39] H. Grabner and H. Bischof. On-line boosting and vision. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 260–267, USA, 2006.

[40] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. In *Proc. of European Conf. on Computer Vision*, pages 234–247, Heidelberg, 2008.

[41] D. Hall. Automatic parameter regulation of perceptual systems. *Image and Vision Computing*, 24(8):870–881, 2006.

[42] Z. Han, Q. Ye, and J. Jiao. Online feature evaluation for object tracking using kalman filter. In *Proc. of Int. Conf. on Pattern Recognition*, Tampa, FL, 2008.

[43] M. Harville. Stereo person tracking with adaptive plan-view statistical templates. In *Euro-*

pean Conference on Computer Vision Workshop on Statistical Methods in Video Processing, pages 67–72, Copenhagen, Denmark, 2002.

[44] J. R. Hoffman and R. P. S. Mahler. Multitarget miss distance via optimal assignment. *IEEE Trans. on Systems, Man and Cybernetics. Part A: Sys. Hum*, 34(3):327 – 336, May 2004.

[45] W. Hu, X. Li, G. Tian, S. Maybank, and Z. Zhang. An incremental dpmm-based method for trajectory clustering, modeling, and retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(5):1051–1065, May 2013.

[46] A. Humayun, O. M. Aodha, and G. J. Brostow. Learning to find occlusion regions. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Colorado Springs, June 2011.

[47] International Telecommunication Union. *Objective perceptual multimedia video quality measurement of HDTV for digital cable television in the presence of a full reference*, 2011.

[48] D. Israel. *Data Analysis in Business Research: A Step-By-Step Nonparametric Approach*. SAGE Pub. Pvt. Ltd., 2008.

[49] ITU-T. Subjective video quality assessment methods for multimedia applications. http://videoclarity.com/PDF/T-REC-P.910-199909-IPDF-E1.pdf. 1999.

[50] N. Johnson and D. Hogg. Learning the distribution of object trajectories for event recognition. *Image and Visual Computing*, 14(8):609–615, August 1996.

[51] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(7):1409–1422, 2012.

[52] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(2):319–336, February 2009.

[53] P. Kranen, H. Kremer, T. Jansen, T. Seidl, A. Bifet, G. Holmes, and B. Pfahringer. Clustering performance on evolving data streams: Assessing algorithms and evaluation measures within moa. In *Proc. of IEEE Int. Conf. on Data Mining Workshops*, 2010.

[54] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cehovin, G. Nebehay, G. Fernandez, and T. Vojir. The vot2013 challenge: overview and additional results. In *Computer Vision Winter Workshop*, Czech Republic, February 2014.

[55] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, F. Porikli, L. Cohovin, G. Nebehay, G. Fernandez, T. Vojir, A. Gatt, A. Khajenezhad, A. Salahledin, A. Soltani-Farani, A. Zarezade, A. Petrosino, A. Milton, B. Bozorgtabar, B. Li, C. S. Chan, C. Heng, D. Ward, D. Kearney, D. Monekosso, H. C. Karaimer, H. R. Rabiee, J. Zhu, J. Gao, J. Xiao, J. Zhang, J. Xing, K. Huang, K. Lebeda, L. Cao, M. E. Maresca, M. K. Lim, M. ELHelw, M. Felsberg, P. Remagnino, R. Bowden, R. Goecke, R. Stolkin, S. Y. Lim, S. Maher, S. Poullot, S. Wong, S. Satoh, W. Chen, W. Hu, X. Zhang, Y. Li, and Z. Niu. The visual object tracking vot2013 challenge results. Technical report, VOT Challenge 2013.

[56] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[57] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39(5):489–504, 2009.

[58] I. Leichter and E. Krupka. Monotonicity and error type differentiability in performance measures for target detection and tracking in video. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[59] H. Li, C. Shen, and Q. Shi. Real-time visual tracking using sparse representation. *http://arxiv.org/abs/1012.2603*, 2010.

[60] H. Li, C. Shen, and Q. Shi. Real-time visual tracking using compressive sensing. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1305–1312, 2011.

[61] T. List, J. Bins, J. Vazquez, and R. B. Fisher. Performance evaluating the evaluator. In *Proc. of IEEE Int. Workshop on Performance Evaluation of Tracking and Surveillance*, 2005.

[62] E. Maggio and A. Cavallaro. *Video tracking: theory and practice*. Wiley, 2011.

[63] D. Makris and T. Ellis. Path detection in video surveillance. *Image and Vision Computing*, 20(12):895–903, October 2002.

[64] D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *IEEE Trans. on Systems, Man and Cybernetics - Part B*, 35(3):397–408, June 2005.

[65] R. Martin and J. M. Martinez. Correlation study of video object trackers evaluation metrics. *Electronics Letters*, 50(5):361–363, 2014.

[66] A. Mayache, T. Eude, and H. Cherifi. A comparison of image quality models and metrics based on human visual sensitivity. In *Proc. of Int. Conf. on Image Processing*, 1998.

[67] X. Mei, H. Ling, Y. Wu, E. Blasch, and L. Bai. Minimum error bounded efficient l_1 tracker with occlusion detection. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2011.

[68] K. Moder. Alternatives to f-test in one way anova in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling*, 52(4):343–353, 2010.

[69] C. Motamed. Motion detection and tracking using belief indicators for an automatic visual-surveillance system. *Image and Vision Computing*, 24(11):1192–1201, 2006.

[70] J. Munkres. Algorithms for assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, March 1957.

[71] C. J. Needham and R. D. Boyle. Tracking multiple sports players through occlusion, congestion and scale. In *British Machine Vision Conference*, pages 93–102, Manchester, UK, 2001.

[72] C.J. Needham and R.D. Boyle. Performance evaluation metrics and statistics for positional tracker evaluation. In *Proc. of Int. Conf. on Computer Vision Systems*, 2003.

[73] A. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. Etiseo, performance evaluation for video surveillance systems. In *Proc. of the IEEE Conf. on Advanced Video and Signal Based Surveillance*, pages 476–481, London, September 2007.

[74] A. T. Nghiem, F. Bremond, M. Thonnat, and R. Ma. New evaluation approach for video processing algorithms. In *IEEE Workshop on Motion and Video Computing*, Austin, Texas, 2007.

[75] J. Ning, L. Zhang, D. Zhang, and C. Wu. Robust mean-shift tracking with corrected background-weighted histogram. *IET Computer Vision*, 6(1):62–69, January 2012.

[76] P. Pan, F. Porikli, and D. Schonfeld. A new method for tracking performance evaluation based on a reflective model and perturbation analysis. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, April 2009.

[77] Y. Pang and H. Ling. Finding the best from the second bests - inhibiting subjective bias in evaluation of visual tracking algorithms. In *Proc. of Int. Conf. on Computer Vision*, 2013.

[78] P. Perez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *Proc. of European Conf. on Computer Vision*, pages 661–675, 2002.

[79] G. Phadke and R. Velmurgan. Illumination invariant mean-shift tracking. In *Proc. of IEEE Workshop on Applications of Computer Vision*, Clearwater Beach, FL, USA, 2013.

[80] C. Piciarelli, G. L. Foresti, and L. Snidaro. Trajectory clustering and its applications for video surveillance. In *Proc. of IEEE Conf. on Advanced Video and Signal Based Surveillance*, Como, September 2005.

[81] H. Pirsiavash, D. Ramanan, and C.C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1201–1208, Colorado Springs, USA, Jun. 2011.

[82] H. Pirsiavash, C. Vondrick, and A. Torralba. Assessing the quality of actions. In *Proc. of European Conf. on Computer Vision*, 2014.

[83] F. Poiesi, R. Mazzon, and A. Cavallaro. Multi-target tracking on confidence maps: an application to people tracking. *Computer Vision and Image Understanding*, 117(10):1257–1272, Oct. 2013.

[84] J. Popoola and A. Amer. Performance evaluation for tracking algorithms using object labels. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, April 2008.

[85] D. Poullin and M. Flecheux. Passive 3D tracking of low altitude targets using DVB (SFN Broadcasters). *IEEE Aerospace and Electronic Systems Magazine*, 27(11):36–41, November 2012.

[86] B. Ristic, B.-N. Vo, D. Clark, and B.-T. Vo. A metric for performance evaluation of multi-target tracking algorithms. *IEEE Trans. on Signal Processing*, 59(7):3452–3457, July 2011.

[87] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision, Special Issue: Learning for Vision*, 2007.

[88] I. Saleemi, K. Shafique, and M. Shah. Probabilistic modeling of scene dynamics for applications in visual surveillance. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(8):1472–1485, August 2009.

[89] S. Salti, A. Cavallaro, and L. D. Stefano. Adaptive appearance modeling for video tracking: Survey and evaluation. *IEEE Trans. on Image Processing*, 21(10):4334–4348, October 2012.

[90] J. C. SanMiguel, A. Cavallaro, and J. M. Martinez. Standalone evaluation of deterministic video tracking. In *IEEE Int. Conf. on Image Processing*, Orlando, 2012.

[91] J.C. SanMiguel, A. Cavallaro, and J.M. Martinez. Adaptive on-line performance evaluation of video trackers. *IEEE Trans. on Image Processing*, 21(5):2812–2823, May 2012.

[92] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/2/3):7–42, 2002.

[93] R. Schubert, H. Kloeden, G. Wanielik, and S. Kaelberer. Performance evaluation of multiple target tracking in the absence of reference data. In *Int. Conf. on Information Fusion*, Edinburgh, July 2010.

[94] D. Schuhmacher, B.-T. Vo, and B.-N. Vo. A consistent metric for performance evaluation of multi-object filters. *IEEE Trans. on Signal Processing*, 56:8, 2008.

[95] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance models for occlusion handling. In *Proc. of Workshop on Performance Evaluation of Tracking and Surveillance*, Hawaii, December 2001.

[96] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance models for occlusion handling. *Image and Vision Computing*, 24(2006):1233–1243, 2006.

[97] R. Sharma, V. I. Pavlovic, and T. S. Huang. Toward multimodal human-computer interface. *Proceedings of the IEEE*, 86(5):853–869, 1998.

[98] S. Stalder, H. Grabner, and L. van Gool. Beyond semi-supervised tracking: Tracking should be as simple as detection, but not simpler than recognition. In *Proc. of Int. Conf. on Computer Vision*, USA, 2009.

[99] X. Suau, J. Ruiz-Hidalgo, and J. R. Casas. Real-time head and hand tracking based on 2.5D data. *IEEE Trans. on Multimedia*, 14(3):575–585, June 2012.

[100] H.-I. Suk, A.K. Jain, and S.-W. Lee. A network of dynamic probabilistic models for human interaction analysis. *IEEE Trans. on Circuits and Systems for Video Technology*, 21:932–945, 2011.

[101] G. Sundaramoorthi, A. Mennucci, S. Soatto, and A. Yezzi. Tracking deforming objects by filtering and prediction in the space of curves. In *Proc. of IEEE Int. Conf. on Decision and Control*, pages 2395–2401, Shanghai, China, Dec. 2009.

[102] C. Tomasi and T. Kanade. Detection and tracking of point features. *International Journal of Computer Vision*, 1991.

[103] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, Apr. 1991.

[104] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013.

[105] C. Vondrick and D. Ramanan. Video annotation and tracking with active learning. In *Proc. of Neural Information Processing Systems*, 2011.

[106] A. Waibel, R. Stiefelhagen, R. Carlson, J. Casas, J. Kleindienst, L. Lamel, O. Lanz, D. Mostefa, M. Omologo, F. Pianesi, L. Polymenakos, G. Potamianos, J. Soldatos, G. Sutschet, and J. Terken. *Handbook of Ambient Intelligence and Smart Environments*, chapter Computers in the Human Interaction Loop, pages 1071–1116. Springer, 2010.

[107] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(3):539–555, March 2009.

[108] X. Wang, K. Tieu, and W. E. L. Grimson. Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(1):56–71, January 2010.

[109] B. L. Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3-4):330–336, 1951.

[110] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele. Monocular visual scene understanding: Understanding multi-object traffic scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(4):882–897, April 2013.

[111] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, Nov. 2007.

[112] H. Wu, A. C. Sankaranarayanan, and R. Chellappa. In situ evaluation of tracking algorithms using time reversed chains. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, 2007.

[113] H. Wu, A. C. Sankaranarayanan, and R. Chellappa. Online empirical evaluation of tracking algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(8):1443–1458, August 2010.

[114] H. Wu and Q. Zheng. Self-evaluation of visual tracking systems. In *Proc. of Army Science Conference*, Orlando, 2004.

[115] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[116] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2034–2041, Providence, Rhode Island, USA, Jun. 2012.

[117] F. Yin, D. Makris, and S. A. Velastin. Performance evaluation of object tracking algorithms. In *Proc. of Workshop on Performance Evaluation of Tracking and Surveillance*, Rio de Janeiro, Brazil, 2007.

[118] K. Zhang, L. Zhang, and M.-H. Yang. Real-time compressive tracking. In *Proc. of European Conf. on Computer Vision*, Florence, Italy, October 2012.

[119] B. Zhou, X. Wang, and X. Tang. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[120] Z. Zhou, X. Chen, Y.-C. Chung, Z. He, T. X. Han, and J. M. Keller. Activity analysis, summarization, and visualization for indoor human activity monitoring. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(11):1489–1498, 2008.